UNIVERSITÄT
DES
SAARLANDES

Fachrichtung 4.7
Allgemeine Linguistik
Philosophische Fakultät II
Universität des Saarlandes
Saarbrücken

Contextually Appropriate Intonation of Clarification Requests in Situated Human–Robot Dialogue

Master Thesis
in Language Science and Technology

vorgelegt von

Raveesh Meena

Angefertigt unter Leitung von
Dr. ing. Ivana Kruijff-Korbayová
und
Prof. Dr. Hans Uszkoreit

July 2010

Meena, Raveesh
Contextually Appropriate Intonation of Clarification Requests in Human–Robot Dialogue
Master Thesis,
Universität des Saarlandes, Saarbrücken, Germany
July 2010 , 150 pages

To my loving grandmother, my grandfather, my first cousin Rajesh and my beloved aunt Jhuma, who all passed away while I was working on this thesis.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Saarbrücken, den 5. Juli 2010

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, July 5, 2010

Raveesh Meena

Abstract

In this thesis we develop an approach for determining contextually appropriate intonation of clarification statements raised during continuous and cross-modal learning in autonomous robots.

Autonomous robots which self-understand and self-extend in the environment in which they find themselves learn continuously about their surroundings. During the course of learning a robot might require additional information from its human interlocutor. Spoken dialogue is a means through which robots can ask their interlocutor for new information, and also for clarifying the knowledge they have acquired about the situated environment.

The ability to self-initiate a dialogue, besides adding autonomy to a robot's behavior, also allows the robot to connect its *belief state* to that of its listener. This enables the participating agents to perform *grounding*, and arrive at a *common ground*. A robot's grounding *feedback* is one of the means to arrive at a common ground. When a robot uses a grounding feedback (e.g. a clarification request) in a given context, it is important for it to be clear how the utterance relates to the preceding context and what it focuses on. Intonation is one means to indicate this relation to context. The task of making the grounding feedback utterances of conversational robots contextually appropriate therefore, inevitably also involves intonation assignment.

Following the analysis of Purver *et al.* [2003] on the forms of clarifications in human dialogue, we develop strategies for formulating clarification requests in human-robot dialogue. The form of a clarification request, its content, and its intonation are all strongly influenced by current contextual details. We represent these contextual factors, communicative intentions, and the corresponding utterance meanings at all levels of processing, in an ontologically rich, relational structures based on Hybrid-Logic Dependency Semantics (HLDS).

As for intonation, we combine the approaches of Steedman [2000a], Lambrecht [1994] and Engdahl [2006] to intonation assignment based on *information structure* (IS), an underlying partitioning of utterance content that reflects its relation to discourse context. The IS units are represented within the same HLDS structure. To achieve prosodic realization from the same grammar as used for utterance realization we extend our OpenCCG grammar for prosody. Following Pierrehumbert and Hirschberg [1990] model of combinatory intonation, we add categories for pitch accents and boundary tones in our grammar. The best realizations, in terms of contextual appropriateness of utterance content as well as its intonation contour, are then post-processed to MaryXML format. This format is finally fed to the MARY text to speech synthesizer for production.

For empirical verification of this approach, we set up psycholinguist experiments to see whether differences in the placement of the main accent in clarification requests are perceivable in synthesized speech, and whether the situated context licenses these accent placement. The preliminary analysis of the data provide evidence for subject's preference of accent placements that are congruent to the visual

context than those that are not congruent to the visual scene.

Acknowledgements

The finishing of this thesis also culminates my master studies in Saarbrücken. Both my studies and working on this thesis have been a wonderful learning experience for me. I would like to take this opportunity to thank all those who have contributed to it, in one way or another.

My first thanks go to my dear friends Christian Ferstl and Gesine Hoinka. If I had not come to know them, I would not have conceived the idea of pursuing higher studies abroad. If it has not been their undying support and encouragement to me in following my dreams, it would not have been possible for me to make it to Germany.

My next thanks go to my supervisors, Prof. Hans Uszkoreit and Dr. ing. Ivana Kruijff-Korbayová. Their advice and guidance have been invaluable to me. I wish to express my particular gratitude to Ivana, with whom I have the privilege of learning to do research. During this work she has been more of a companion, right from the selection of the topic to identifying the relevant literature; from the numerous subsequent discussion on intonation and information structure to designing experiments and analyzing results. Thank you for your constant support and encouragement, and for your patience in reading and commenting on the successive draft versions.

I would also like to express my thanks to our project leader at DFKI, Dr. ir. Geert-Jan M. Kruijff, for introducing me to the research domain of Human-Robot Interaction. He has been instrumental in shaping the course of this research work from the very beginning till the end. If his name is not on the cover page, then it is only due to administrative reasons.

I am thankful to Pirita Pyykkönen for providing us with her expertise in setting up the psycholinguistic experiments and analyzing the data.

I am also thankful to the other members of the CoSy/CogX research group at DFKI: Pierre Lison, Miroslav Janíček, Dennis Stachowicz, Christopher Koppermann, Hendrik Zender and Sergio Roa. It has been a great pleasure to work with you.

Many thanks to my friends Verena Stein, Eva Wollrab, Daniel Bauer and William Roberts, for being home to me in this tiny yet lovely "city". It has been a really beautiful time that I spent with you.

Last but not least, I would like to thank my family for their love and constant support. My apologies to my parents for whom these two years have been an emotional ordeal due to the physical distances between us. I'm indebted to my sisters Smita and Priyanka, for their unconditional support to me and my aspirations, and for being *the responsible sons* during my absence at home.

Raveesh Meena
Saarbrücken
July 5, 2010

Table of contents

1 Introduction

In this introductory chapter we give an overview of the thesis. We start with describing the problem and the research domain to which this work pertains. During this, we will draw an outline of the main research questions pursued in this thesis. Next, we come to the main claim of this thesis and discuss how we address these research questions. Following this we describe the application platform in which we develop this research work. Towards the end we discuss the state-of-the-art in the problem domain and state the contributions of our research work to it. We close with an outline of the remaining parts of this thesis.

1.1 The Problem

Recent years have witnessed a trend towards developing a new generation of robots that are capable of moving and acting in human-centered environment. These robots interact with people and participate in our everyday life activities. As assistive partners they help humans in daily chores on a shared basis or even autonomously. An essential characteristic of these autonomous robots is their ability to continuously learn about their surroundings.

Continuous learning

Continuous learning requires a robot to be able to *self-understand* and *self-extend*. What this means is that the robot has an understanding of what it knows and does not know about the world it finds itself in. And when the robot finds out that there is something it doesn't know or is uncertain about, it is able to plan actions to seek information to fill these knowledge gaps or for clarifying the uncertainties. By planning actions thus a robot can not only acquire new information about the world but also new skills to enhance its abilities.

spoken dialogue

Among these possible actions, initiating a *spoken dialogue* with its human partner is a means through which a robot can request information or clarify its doubts about the surroundings. For this to work, the robot and the human need to first establish a mutually agreed-upon understanding of what is being talked about, and why - that is, they need to reach a *common ground*. Especially when asking information questions or requesting clarifications, the robot needs to indicate very clearly the objects and their properties it is after.

For example, consider the scenario in Figure 1.1, where the robot is trying to automatically learn about the properties of the object lying on the table. Here,

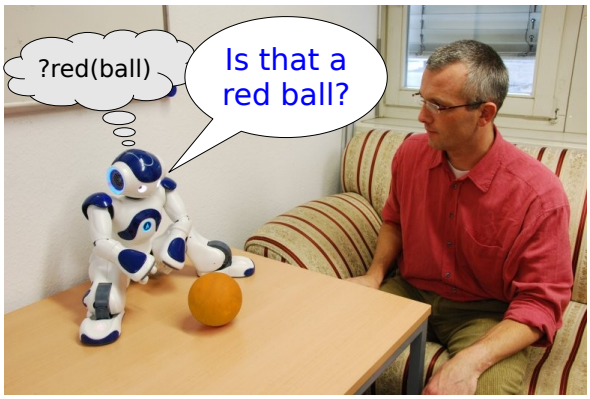


Figure 1.1: Requesting Clarification

it appears that the robot is certain that the object is of type 'ball', however, it is not certain if this object also has the color 'red'. Suppose that the robot comes up with a clarification request "*Is that a red ball?*" to resolve this *uncertainty* about the color. Now, intonation plays an important role here. The sentences in (1) illustrate two different intonation contours for this clarification request (CR). The words in SMALLCAPS indicate the alignment of the main pitch accent stress in the intonational contour of the utterance.

intonation

(1) a. Is that a red BALL?

b. Is that a RED ball?

Intuitively, and in line with the existing work on intonation and its role in the interpretation of discourse meaning, the accentuation in these two sentences mark the part of the utterance with the highest need for verification, whereas that assumed to have been correctly recognized, and thus part of the common ground between the robot and human tutor, remain unaccented.

Accordingly, the robot's intentions in the current scenario are expressed more clearly by the clarification request in (1b). The clarification request in (1a) on the other hand is more appropriate when the robot assumes the color to be correctly recognized but is uncertain about the type of the object on the table.

Therefore when requesting clarification, in order to be able to convey the intended meaning to its human partner, a robot needs to be able to produce the utterance with the appropriate intonation. This requires the robot to have a deliberate control over the intonation of the utterance being produced.

The need of deliberate control over the intonation of an utterance is not limited to just clarification requests and information questions. A robot needs to be able to also control the intonational contours of its responses if it were to contribute meaningfully to a dialogue. For example, consider the scenarios in Figure 1.2 where the robot has been asked about the color of the 'box', and in Figure 1.3 where the

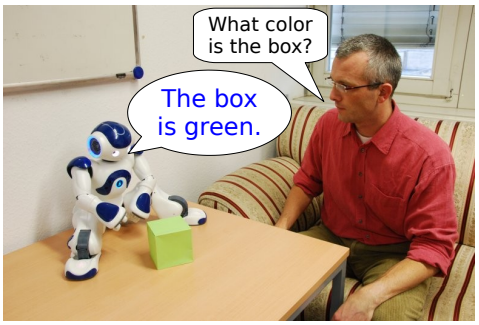


Figure 1.2: Responding to a human query about object *color*.

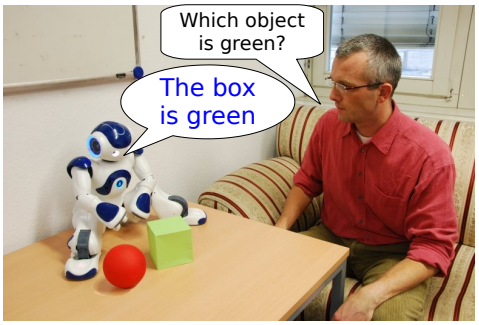


Figure 1.3: Responding to a human query about object *type*.

questions is about the type of the 'green' object.

Observe that in both the scenarios the robot's response[1] contains the same sequence of words, as shown in (2) and (3), respectively. Also, in (3b) the accentuation on the color 'green' is mandated by the *situated context* in the scenario of Figure 1.3, where the presence of the object 'ball' and thereof the color 'red', necessitates a distinctive contrast on 'green'.

situated context

(2) a. H: What color is the box?
b. R: The box is GREEN.

(3) a. H: Which object is green?
b. R: The BOX is GREEN.

These responses differ in the *intention* they carry, in what they indicate about the robot's *attentional state* and the contextual aspects of the dialogue.

intention

attentional state

It is evident from these scenarios that it is critical for a conversational robot to have a deliberate control over the intonation of its utterances. Moreover, the assignment of intonation to an utterance needs to reflect upon the robot's intentional and the attentional state with respect to its beliefs, in a dynamic dialogue and visual context. How to assign contextually appropriate intonation contour to a robot's utterances in a situated dialogue is the fundamental research question that we pursue in this thesis work.

1.2 The Claim

In pursuit of this research objective, we claim that: *The contextually appropriate intonational realization of robot utterances such as questions, clarifications requests*

[1]Throughout this work we use the notation H and R to indicate human and robot utterances respectively in a dialogue.

and responses can be established through the interplay of intention and intension, relative to a robot's belief models.

common ground

As a starting point in this direction, it is important to bear in mind that the function of an utterance in a dialogue is to establish (and extend) a *common ground*, provide new information that extends or otherwise modifies information that becomes shared [Clark and Schaefer, 1989]. An utterance *reflects* the speaker's (S) cognitive state as to what S believes, what S intends, what it knows and does not know, and also what it believes and presumes about the hearer (H). At the same time, the utterance *affects* H's cognitive state, as to what H believes, what H intends and plans to do next, and what H knows about S. The notion of *information structure* in an utterance is a presentational means which the speaker employs to achieve the contextually appropriate realization of the information that is being communicated.

information structure

The information structure (IS) of an utterance is an underlying partitioning of the utterance content that reflects its relation to dialogue context. It indicates how an utterance links to the preceding dialogue – what has happened or has been talked about so far, and also what the utterance contributes to the current dialogue. In spoken English, the information structure of an utterance is predominantly realized by its intonation contour [Steedman, 2000a].

Following this, the information structure partitioning of the robot utterances in (2) and (3) is indicated in (4) and (5) respectively. The brackets with subscript Th mark the contents of an utterance which link it to the preceding dialogue (also referred to as the *theme*). On the other hand the brackets with subscript Rh mark the contents which contributes additional information to the current discourse (also referred to as the *rheme*).

(4) a. H: What color is the box?
b. R: (The box is $)_{Th}$ (GREEN $)_{Rh}$

(5) a. H: Which object is green?
b. R: (The BOX $)_{Rh}$ (is GREEN $)_{Th}$.

It is noteworthy to observe that utterances (4b) and (5b) differ in their information structure partitioning, the presence and absence of accentuation in the theme partitions and therefore also in their presentation. These differences at the presentational level are the reflection of the robot's belief state and its intentional and attentional state in the corresponding dialogue context. This illustrates that the task of assigning a contextually appropriate intonation contour to an utterance begins with the assignment of a contextually appropriate information structure to the utterance.

What we have discussed so far outlines the major research objectives which we pursue in this thesis:

Research Goal 1. *Modeling an utterance's information structure assignments relative to the speaker's cognitive state in a dynamic dialogue context.*

Research Goal 2. *Intonational realization of an utterance's information structure.*

To achieve these objectives we:

- base questions and clarification requests in a multi-agent belief model that gives rise to them.
- determine information structure using the model of agent's intention and attention.
- use Steedman [2000a]'s Combinatory Categorical Grammar theory to establish an interface between semantics and prosody.
- provide an extended model to cover more types of utterances with particular focus on clarification requests.

During the course of this thesis we will provide thorough details on each of these aspects of our approach. In the following section we discuss the application scenario in which we develop this research work. Our objective here is to illustrate the type of human-robot dialogue we aim to achieve in this work, and also emphasize the non-trivial role of intonation in a situated dialogue.

1.3 Application Scenario

The work presented in this thesis is being conducted as part of the CogX[2] project – *Cognitive Systems that Self-Understand and Self-Extend* (May 2008 – July 2012). It is a large-scale integrated project funded in the European Union's Seventh Framework (EU FP7), as part of the Cognitive Systems thematic area. The consortium involves six universities and about 30 researchers.

One of the primary research aims of the CogX project is to investigate how a cognitive system could continuously acquire new knowledge and new skills in a life-long manner. Towards this, approaches are being developed[3] for *interactive continuous learning* of visual concepts, in particular, learning object colors and some basic shapes.

George[4] is a scenario that illustrates the developed functionalities. George is one of the robot that is being developed in CogX project. It is capable of visual processing, communication and learning. The dialogue fragment in (6) illustrates some of the interesting aspects of the interactive continuous learning scenario in Figure 1.4 and Figure 1.5 .

(6) a. H: Hi robot.

b. R: Hello.

[2]Official website of CogX project: http://cogx.eu/

[3]at *DFKI GmbH*, Saarbrücken, Germany and *University of Ljubljana*, Slovenia

[4]http://cogx.eu/results/george/

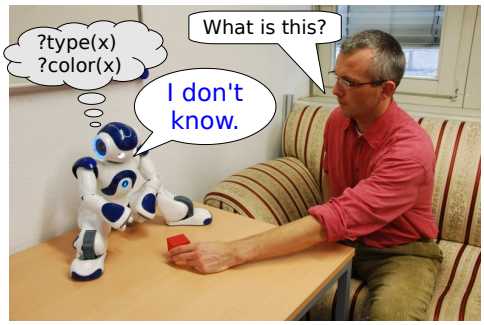


Figure 1.4: Indicating an empty model about concept *type*.

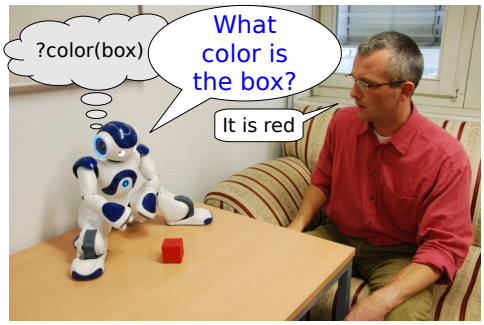


Figure 1.5: Learning model for concept *color* via dialogue.

c. H: *puts a red box in front of the robot*
d. R: *starts low-level visual learning*
e. R: What is that? [Situated Grounding feedback]
f. H: Do you know what this is?
g. R: No, I do not know what this is. [Pre-linguistic learning, Situated Grounding]
h. H: This is a box. [Tutor driven learning]
i. R: Let me see. [Linguistic Grounding feedback].
j. R: I see. [Linguistic Grounding feedback].
k. R: What COLOR is the box? [Situated Grounding feedback]
l. H: It is red.
m. R: Let me see.
n. R: I see.

At the onset the system doesn't have any models of concepts about visual properties like color and type. Therefore despite having made some pre-linguistic detections in step (6d) George cannot recognize the object. George notifies its human partner about this situated reality by the grounding feedback in (6e). George's response in (6g) to the human query in (6f) indicates its lack of knowledge, which prompts the human partner for tutoring.

(6h) indicates a tutor based learning instruction. George uses this information input to build a model for object *type* 'box' by grounding the linguistic input to sensory data i.e. associating them with the features extracted from the low-level sensory data. George provides the feedback for this grounding in (6i). Using the combination of low-level visual learning, and learning based on communication with the tutor, George builds models of concepts and acquires knowledge about its surroundings. George acknowledges that it has successfully built the model for 'box' with its response in (6j).

Curious as he is, George continuously acquires low-level visual features about visual objects in its view and uses these features to update existing models of concepts or build a new model of concept. When George can't build a model for these features, he initiates a dialogue with the human partner to seek model or conceptual information for them. The request for information in (6k) is one such attempt by George to learn the *color* property of the model 'box'. With the human response in (6l) George builds a model for color 'red' by grounding it with the low-level visual features hue. George acknowledges the success of grounding in (6n).

One interesting aspect of George's communicative ability is that of using *grounding feedbacks* as in (6e), (6i) and (6j) which enables the human partner to learn about what is going on in George's head i.e. his cognitive state. George acts on its own understanding, which need not be in any way similar to how its partner sees the world. There is therefore a need for the robot to make clear what it is after: why the robot is requesting something from a human, what aspects of a common ground it appeals to, and how the request is related to what it does and does not know.

The interactive learning scenario in (6) can extend further to the scenarios discussed in previous section (refer Figure 1.1, 1.2 and 1.3), where the contextually appropriate intonation of an utterance is critical for avoiding situational ambiguity in a dialogue. For example, even a rather straightforward utterance such as "*What color is the box?*" may require explicit control of intonation in different situations, as enumerated in (7).

(7) a. What COLOR is the box?
    b. WHAT color is the box?
    c. What color is the BOX?

The realization of this sentence with the accentuation of 'color', as in (7a), indicates the speaker's intention to seeks the 'color' value of the only salient object type 'box'. The same sentence when realized with accentuation of 'what', as in (7b), indicates the speaker's request for clarification over a color value that has been just mentioned. When realized with the accentuation of 'box', as in (7c), the utterance indicates the speaker's intention to seek the color value of a specific 'box' among the set of other salient object types.

The need of contextually appropriate intonation in robot utterances cannot be undermined in a situated dialogue. In the due course of this thesis we will describe our approach to fulfill this demand. In the following section we discuss the state-of-the-art in the research domain of producing contextually appropriate intonation in system generated utterances.

1.4 State-of-the-Art

Intonation has been studied for quite sometime now, both for interpretation the discourse meaning of utterances [Pierrehumbert and Hirschberg, 1990] and for the

realization of the discourse information structure of utterances [Steedman, 2000a]. Almost all of the practical applications that deal with the task of production of intonation in system generated responses are based upon Steedman's theory of information structure [Steedman, 2000b,a].

Following the notions developed in [Pierrehumbert and Hirschberg, 1990], on the contribution of intonation contours in the interpretation of discourse, Steedman proposes a unified theory that draws a relation between context, grammar and intonation as a reflection of the information structure partitioning of an utterance.

Some of the practical systems which employ Steedman's theory of IS for intonation realsation include [Prevost, 1996], [Kruijff-Korbayová *et al.*, 2003], [White *et al.*, 2004b] and [White *et al.*, 2004a]. The wide appeal of Steedman's theory of IS and the approach to combinatory prosody is due to the fact that it:

- associates intonation with discourse meaning in terms of information structure;
- provides a compositional semantics of English intonation in information-structural terms;
- tightly couples intonation with grammatical structure;
- assumes a general IS-sensitive notion of discourse context update;
- has proved its worth in previous practical applications to control intonation assignment with respect to context.

Early work on controlling the intonation of synthesized speech with respect to context concerned mainly accenting open-class items on first mention, and deaccenting previously mentioned or otherwise "given" items [Hirschberg, 1993; Monaghan, 1994]. But such algorithms based on givenness fail to account for certain accentuation patterns, such as marking explicit contrast among salient items. Givenness alone also does not seem sufficient to motivate accent type variation.

In [Prevost, 1996] contrastive accent patterns and some accent type variation are modeled using Steedman's approach to IS in English. In one application he handles question-answer pairs where the question intonation analysis in IS terms is used to motivate the IS of the corresponding answer, realized through intonation. Another application concerns intonation in generation of short descriptions of objects, where Theme/Rheme IS partitioning is motivated on text progression grounds, and Background/Focus IS partitioning distinguishes between alternatives in context.

In [Kruijff-Korbayová *et al.*, 2003], [White *et al.*, 2004b] and [White *et al.*, 2004a], a similar IS-based approach is applied to assign contextually appropriate intonation to the output of actual end-to-end dialogue systems (German and English, respectively). The reported evaluation results show that this leads to qualitative improvements.

The intonation of questions, and CRs in particular, has so far been largely neglected in dialogue systems. The practical applications mentioned above all concentrated on the assignment of intonation in statements. However, a series of production and perception experiments around the HIGGINS dialogue system [Edlund *et al.*, 2004], show that fragmentary grounding utterances in Swedish differ in prosodic features depending on their meaning (acknowledgment vs. clarification of understanding or perception), and that subjects differentiate between the meanings accordingly, and respond differently [Edlund *et al.*, 2005; Skanze *et al.*, 2006].

In a study of a corpus of German task-oriented human-human dialog, [Rodríguez and Schlangen, 2004] also found that the use of intonation seemed to disambiguate clarification types, with rising boundary tones used more often to clarify acoustic problems than to clarify reference resolution.

1.5 Contributions of the thesis

- Our work extends the use of information structure to control the intonation of dialogue system output beyond answers to *information-seeking questions*, *clarification requests* and *acknowledgements*.
- We include both fragmentary *grounding feedback* and full utterances, and address varying placement of pitch accents depending on context and communicative intention.
- Our approach focuses specifically on gathering the contextual details in a dynamic discourse context, which comprises agent beliefs about the *visual scene* and the *dialogue* history.
- The novelty in our methods is how they achieve to flexibly combine *intention*, *multi-agent beliefs*, and *attentional state* in continual processing of dialogue.
- We present an *implementation* for realizing contextually appropriate intonation of robot utterances in a situated human-robot dialogue.
- We present an implementation of *combinatory prosody* in OPENCCG platform.
- We present a psycholinguistic experiment for the *verification* of the contribution of *visual context* in determining the contextual appropriateness of intonation contour in robot utterances in a situated dialogue.

1.6 Outline of the thesis

This thesis is organized into three main parts: Background, Approach, and Verification & Conclusion.

Background In the Background part we discuss the literature and the necessary theoretical background for the development of this research work. In Chapter 2, we discuss the pioneering work of Pierrehumbert and Hirschberg [1990] on the contribution of intonational contours in interpretation of discourse meaning, and highlight its relevance to our work. After this, we briefly touch upon some of the important theories on the formulation of discourse information structure, and describe Steedman [2000a]'s theory of IS, which is the foundation on which we base our research work.

In Chapter 3, we start with an overview of the CogX system architecture, and show where our research work fits in the pipeline. Then we discuss Hybrid Logic Dependency Semantics, which is the means of knowledge representation in the system. Next, we discuss the belief models and associated inference methods which are at the core of our dialogue system. Towards the end we discuss the approach to utterance content planning, and the grammar framework for utterance and intonation realization, namely Combinatory Category Grammar (CCG).

The Approach In this part we discuss our approach to modeling information structure in robot utterances and its intonational realization. In Chapter 4, we lay down the general principles for a IS based presentation of robot's utterances using agent's beliefs, intention and attentional state. We present an implementation of IS assigning to robot utterances. In Chapter 5, we present our approach to model Steedman's combinatory grammar in the OpenCCG platform. We illustrate the modeling of the inventory of a prosodic grammar, and elaborate our approach to model prosodic derivations. We end the chapter with a discussion on the theoretical formulation vs. prosodic realization of IS.

Verification & Conclusion In Chapter 6, we describe our approach to experimental verification of the central claim made in this thesis. We start by briefly introducing various schemes for evaluating/measuring this work. Following this, we motivate the chosen methodology for the ongoing experiment. We elaborate on the experimental setup, the parameters, the design and the procedure. We conclude with a discussion on our findings and directions for future work. Chapter 7 concludes our thesis. We present a summary of what we have achieved in this work. We discuss the findings of the ongoing experimental verification of our approach, and outline plans for further investigations. We then provide suggestions for further research on this work.

Part I

Background

2

Intonation and Information Structure

In this chapter, we give an overview of the background literature relevant to the realization of intonation in utterances. We start by reviewing the investigations of Pierrehumbert and Hirschberg [1990] on the contribution of intonation to utterances and discourse interpretation. We elaborate on their compositional theory of tune interpretation. Next, we briefly touch upon some of the more successful theories of information structure, which provide for intonational realization and interpretation of utterance meaning. We deliberate on Steedman [2000a]'s theory of information structure which is the foundation on which we base the research carried out in this thesis. We end with a brief discussion on the compositional theories of intonation proposed by Steedman and Pierrehumbert and Hirschberg [1990] respectively.

2.1 Meaning of Intonational Contours

In examining the particular contribution of the choice of tune, or *intonational contour*, to discourse interpretation, Pierrehumbert and Hirschberg [1990] propose that a speaker (S) chooses a particular tune to convey a particular relationship between the utterance, currently perceived *beliefs* of a hearer or hearers (H), and anticipated contributions of subsequent utterances. These relationships are conveyed compositionally via selection of *pitch accents*, *phrase accents*, and *boundary tones* that make up tunes.

In this section we review the key aspects of the compositional theory of tune interpretation presented by Pierrehumbert and Hirschberg [1990] (now onwards P&H). The notations used here for intonational description follows from [Pierrehumbert, 1980; Pierrehumbert and Beckman, 1986]. Almost all of the examples used in here for illustrations are taken from P&H.

In describing intonational patterns P&H distinguish *stress*, *tune*, *phrasing* and *pitch range* as the dimensions along which intonational variation takes place. The *stress* pattern of an utterance is the pattern of relative prominence of the syllables. *Word stress* is assigned by lexical-phonological rules. Stress assignment within the phrase is influenced by the considerations of *information structure*. For example,

stress

the following sentence would usually be pronounced with the main phrasal stress (the *nuclear stress*) on the word *vitamins*:

(8) Legumes are a good source of VITAMINS.

However, the nuclear stress would fall on *good* in a context where *sources of vitamins* are already under discussion (i.e. *given*), as in (9):

(9) A: Legumes are a pretty poor source of vitamins.
B: No. Legumes are a GOOD source of vitamins.

In an utterance, syllables with greater stress are more fully articulated then syllables with less stress.

tune

Tune is the abstract source of fundamental frequency patterns and is described as a sequence of *low* (L) and *high* (H) tones, which determine the shape of the f_0 *contour*. Some of these tones go with stressed syllables. The other tones, the *phrasal tones* mark the edges of phonological phrases.

pitch accents

Pitch accent mark the lexical items with which they are associated as prominent. [Pierrehumbert and Beckman, 1986] identify six different types of pitch accent in English. These include two simple – high and low – and four complex ones. The high tones, the most frequently used accent, comes out as a peak on the accented syllable. It is represented as H*. The "H" indicates a high tone and the "*" that the tone is aligned with a stressed syllable. Tone L* occur much lower in pitch range and are phonetically realized as f_0 minimas. The other tones are L+H*,L*+H, H*+L, and H+L* where the "*" indicate the alignment of the tone with the stressed syllable.

phrasing

Pierrehumbert and Beckman [1986] report that two levels of *phrasing* in English are involved in the specification of tune. These are *intermediate phrase* and *intonational phrase*. A well-formed intermediate phrase consists of one or more pitch accents, plus a simple high or low *phrase accent* (a H or L tone), which marks the end of the phrase. A phrase accent controls the f_0 between the last pitch accent of the intermediate phrase and the beginning of the next intermediate phrase – or the end of the utterance.

Intonational phrases are composed of one or more intermediate phrases. The end of intonational phrase is marked by additional H or L tones, which are referred to as *boundary tones* and are indicated with a diacritic "%" to distinguish them from phrasal accents. The tones falls exactly at the phrasal boundary. Since the end of every intonational phrase is also the end of an intermediate phrase there are altogether four ways that a tune can go after the last pitch accent of an intonational phrase: LL%, HL%, LH%, HH%.

A phrase's tune or *melody* is defined by its particular sequence of pitch accent(s), phrase accent(s) and boundary tone. For example, an ordinary declarative pattern with a final fall is represented as H* L L%, a tune with H* pitch accent, L phrase accent and L% boundary tone.

pitch range

Another dimension of variation in a tune is the *pitch range*. When a speaker's voice is raised, the overall pitch range – the distance between the highest point f_0 contour and the *baseline* – is expanded. *Final-lowering* is a local-time dependent

final-lowering

type of pitch range variation associated with declaratives, where the pitch range is lowered and compressed in anticipation of the end of the utterance.

Both overall pitch range and final lowering affect the interpretation of a intonational tune. They contribute to the hierarchical segmentation of the discourse. For example, it has been observed that the final lowering reflects the degree of "finality" of an utterance; the more final lowering, the more the sense that an utterance completes a topic. In addition to its role in signaling the overall discourse structure, pitch range interact with the basic meanings of tunes to give their interpretations in context.

2.1.1 A Compositional Approach to Tune Meaning

In the literature tunes have been portrayed as conveying speaker attitude, emotion, speech acts, propositional attitudes, presupposition, focus of attention etc., however, only a few of these characterizations have been successful for particular tunes, and none seems appropriate as a general approach.

Though speaker attitude may sometimes be inferred from the choice of a particular tune, the many-to-one mapping between attitudes and tunes suggests that attitude is better understood as derived from the tune meaning interpreted in *context* than as representing that meaning itself. context

On the basis of the individual tunes that have been studied, P&H argue further that tune meaning is more usefully viewed as *compositional*. They propose that tunes that share certain tonal features seem intuitively to share some aspects of meaning. For example, various types of question contours, L* H H% and H* H H% do share common high phrase accents and boundary tones while differing in the pitch accents used with them.

In their compositional approach to tune meaning P&H propose that a speaker S employs a tune to modify what (S believes) a hearer H believes to be *mutually believed*. This could be S's use of tune in terms of the *intention* to add to what (S believes) H believes to be mutually believed –or not– or to call *attention* to certain relationship between propositions realized by an utterance or other mutually believed propositions. intention attention

P&H suggest that aspects of *intentional structure* as well as the *attentional structure* of a discourse can be conveyed by choice of tunes. For example, S may seek to inform H of some proposition x by communicating that x is to be added to what H believes to be mutually believed between S and H –via the tune S chooses. And S may seek to convey the information status of some item y –say, that y is old information that is to be treated as particularly salient – by the type of accent S uses in realizing y. S's beliefs are however *not* specified by choice of tune –the "declarative" contour H* L L%, for example, will not be translated as "S *believes* x". But S's belief in x may be inferred from the combined meaning of pitch accents, phrase accents, and boundary tones, as they are used in particular contexts.

As per their notion of compositional intonation, pitch accents, phrasal tones and boundary tones each operate on a (progressively higher) domain of interpretation.

Each level contributes a distinct type of information to the overall interpretation of a tune.

Pitch accents convey the information status of discourse referents, modifiers, predicates, and relationships specified by the lexical items with which the accents are associated. Accenting or deaccenting of items in general is associated with S's desire to indicate the relative *salience* of accented items in the discourse. Accent type indicates whether items or predications are to be added or excluded from mutual beliefs that H holds, whether something predicated of these items should be inferable from beliefs H already holds, or whether relationships in which S believes the items participate should be identified by H.

salience

For example, each H* accent in (10) provides information that S intends H to add the marked items to their mutual beliefs.

(10) The train leaves at seven.
H* H* H* L L%

The phrase accents convey information at the level of intermediate phrase. In (10) there is but a single intermediate phrase, marked with a L phrase accent. In (11), however, there are two:

(11) The train leaves at seven or nine twenty-five.
H* H* H* H H* H* L L%

P&H propose that S chooses phrase accent type to convey the degree of relatedness of a phrase to the preceding and succeeding intermediate phrase(s). When the phrase *the train leaves at seven* has a H phrasal accent, for example in (11), it is more likely to be interpreted as a unit with a phrase that follows. The L phrase accent on the other hand doesn't convey any such relation, as can be observed in (10).

Boundary tones contribute information about the intonational phrase as a whole. They convey information about relationships among intonational phrases – in particular about whether the current phrase is to be interpreted with particular respect to a succeeding phrase or not. For example, in (12), S can indicate by a H boundary tone in (12a) that (12a) is to be interpreted with particular respect to a succeeding phrase (12b). The *forward reference* signaled by the boundary tones in (12) might be interpreted as indicative of a hierarchical relationship.

(12) a. The train leaves at seven.
H* H* H* L H%

b. It'll be on track four.
H* H* L L%

To sum up P&H's compositional theory of tune interpretation, the tune meaning is composed of three types of tones – pitch accents, phrase accents and boundary tones –which have scope over three different domains of interpretation. Together,

these intonational features convey how S intends that H interpret an intonation phrase with respect to (1) what H already believes to be mutually believed, and (2) what S intends to make mutually believed as a result of the current utterance.

2.1.2 The Interpretation of Pitch Accents

All pitch accents render salient the material with which they are associated. The accented material is salient not only phonologically but also from an informational standpoint. If the logical form corresponding to an intonational phrase is viewed as an *open expression* in which the accented items are replaced by variables, then the pitch accent marking in S's utterance indicate the items with which H should instantiate these variables. For example, the utterance in (13) might be represented as a logical form in (14), where the accented items are replaced by variables x and y.

(13) George likes pie.
H* H* L L%

(14) x likes y
x(H*)
y(H*)
x = George, y = pie

The S's instantiation of the accent bearing variables x and y with *George* and *pie* respectively, indicates S's intention that it wants the H to instantiate these variables with specific propositional values (and not any other) and add them to their mutual beliefs.

The H* and L* Accents

The H* accent conveys that an item made salient by H* is to be treated as *new* in the current discourse. Stated otherwise, a H* accent appears to signal to the hearer H that the open expression is to be instantiated by the accented items and the instantiated proposition realized by the phrase is to be added to H's mutual belief space.

The combination of H* with a L phrasal accent and a L or H boundary tone i.e. H* L L% is the contour for "neutral declarative intonation". This is appropriate when S's goal is to convey information, as in (15).

(15) My name is Mark Liberman.
H* H* L L%

On the other hand H* H H% is a contour for *high-rise questions* which is the preferred choice when the question phrase simultaneously conveys information, as in (16), where the speaker provides information about his identity at a reception desk and poses a question to confirm if he is at the right place.

(16) My name is Mark Liberman.
H* H* H H%

Here, the H* accent conveys that information is to be added to H's mutual beliefs, and the H phrase accent and boundary tone "question" the relevance of that information.

The L* accent marks items that S intends to be salient but not to form part of what S is predicting in the utterance. It can be said that S uses L* when it can't make predications about marked entities, which indicates that the S believes the current instantiation of the open expression to be uncertain.

L* accent commonly appears in canonical *yes-no questions* – L* H H%. For example, in (17), by marking *prunes* and *feet* with accent L* the speaker S makes no predications about them, however, S desires that H makes such predication.

(17) Do prunes have feet.
L* L* H H%

S can also use L* when it believes that instantiated expression is part of H's mutual belief. In such cases L* goes with the contour L* L H%, where it plays the role of reminding or reassuring.

P&H argue that there are evidences that L* is also used for extra-positional, such as greetings, vocatives, and so called cur-phrases.

Generally speaking L* accent is used by S to exclude the accented items from the predication S intends to be added to H's mutual beliefs. On the other hand a speakers use of H* accent is viewed in terms of attempted modification of H's mutual belief.

The L+H and H+L Accents

P&H propose that the complex pitch accents like L+H and H+L are employed by S to convey the salience of some *scale*, linking the accented item to other items salient in H's mutual belief. However, with H+L accents S intends to indicate that support for the open expression's instantiation with accented item should be inferable by H, from H's representation of the mutual beliefs. The inference can be direct or indirect, and it can be (often) pragmatic rather then logical in character.

A speaker chooses the L*+H pitch accent to convey lack of predication and to evoke a scale. For example, it has been observed that the interpretation of contour L*+H with L phrase accent and H boundary tone (L*+H L H%) expresses *uncertainty* about a scale evoked in the discourse. For example, in (18), B expresses uncertainty about whether being a good badminton player provides relevant information about the degree of clumsiness:

(18) a. A: Alan's such a klutz.

b. B: He's a good badminton player.
L*+H L H%

On the other hand pitch accent L+H* evokes a salient scale. However, S employes the L+H* accent to convey that the accent item – and not some alternative related item – should be *mutually believed*. This can convey the effect of speaker *commitment* to the instantiation of the open expression with the accented item.

The most common use of L+H* has been observed as marking corrections or *contrast*. Here S substitutes a new scalar value for one previously proposed by S or by H – of for some alternative value available in the context. The fall-rise pattern of L+H* has also been used for associating marked items with "background" information. contrast

(19) a. A: I ate the chips. What about beans? Who ate them?

b. B: Fred ate the beans.
H* L L+H* L H%

In (19b) the L+H* accent has been used to contrast *beans* with the alternative *chips* and also for representing the background, which has been established by (19a).

Like the L+H accents, the H+L accents are used by S to evoke a particular relationship between the accented items and H's mutual beliefs. When using H*+L accent, S appears to be making a prediction in the same sense as when using H*, but differs in conveying that H should locate an inference path supporting the predication. On the other hand, S uses H+L* to convey that the desired instantiation of an open expression is itself among H's mutual beliefs. This is related to L* tone, where S does not make a predication.

With these descriptions of the meaning of the pitch accents, P&H observe that the meaning of starred tones are shared among the different accent types. When the starred tone is L (L*, L*+H, H+L*), S does not convey that the instantiation of the open expression should be added to H's mutual beliefs. However, when the starred tone is H(H*, L+H*, H*+L), S does intend to instantiate the open expression in H's mutual belief space. Tones L*+H and L+H* both evoke a salient scale, and H*+L and H+L* both convey that H should be in a position to infer support for the instantiated expression.

2.1.3 The Interpretation of Phrasal and Boundary Tones

Phrasal accents have scope over entire intermediate phrases and may consist of either a high (H) or a low (L) tone. These tones appear to indicate the presence or absence of an interpretive as well as a phonological boundary. A H phrase accent indicates that the current phrase is to be taken as *forming* part of a larger composite interpretive unit with the following phrase. On the other hand a L phrase accent emphasizes indicates separation of current phrase from the subsequent phrase. The phrase accent usage in (11) illustrate this.

Boundary tones have scope over the entire intonational phrase. They play a considerable role in the indication and the perception of discourse segments. P&H propose that the choice of boundary tone conveys whether the current intonational

phrase is *forward looking* or not – that is whether it is to be interpreted with respect to some succeeding phrase or whether the direction of interpretation is unspecified.

H boundary tone indicates that S wishes H to interpret an utterance with particular attention to subsequent utterance, forward-looking, whereas boundary tone L doesn't indicate such directionality. H% can be interpreted as signaling a hierarchal relationship between intentions underlying the current utterance and a subsequent one. Thus H% helps in marking discourse segments. The "forward reference" purpose of H boundary tone differs from their use in *yes-no* questions where H% is used by S to elicit response e.g. in *who*-questions.

Though P&H's compositional theory of tune interpretations is only a first approximation, it nevertheless brings to light the role of intonational contours in reflecting a speaker's beliefs, intentional and attentional state in a dialogue context. Furthermore, their theory has relevance to the realization of the information structure meaning of an utterance. In the following section we discuss in brief some of the theories of information structure and particularly Steedman [2000a]'s theory, which is the foundation on which we base our current thesis work.

2.2 Discourse and Information Structure

The term *information structure* (IS) goes back to Halliday [1967] and has been widely used in the subsequent literature to refer to the partitioning of an utterance's content into categories such as focus, background, topic, comment, rheme, theme and etc. Related notions include Chafe [1974]'s *information packaging* as well as the *functional sentence perspective* of the Prague school [Firbas, 1975]; [Sgall *et al.*, 1986]. There is, however, no consensus on what and how many categories of information structure should be distinguished, or how these can be identified [Kruijff-Korbayová and Steedman, 2003].

Generally speaking information structure is a means that the speaker employs to present some parts of an utterance's meaning as relating it to the preceding discourse and the other parts as contributing new information to the current context. Depending on the type of language, information structure may be indicated in the surface form of the sentence through a combination of word order, tune, and morphology.

For example, in languages that have relatively free word order, like Czech , information structure is primarily realized through variation in word order. However, intonation still remains an integral part of the realization with the focused word carrying the nuclear accent. On the other hand languages with a rigid word order, like English, predominantly employ tune, punctuation, or (marked) syntactic constructions for IS realization.

Information structure has been studied and developed along various lines of thoughts. Within the Functional Generative Description (FDG) framework of the Prague School, the notion of IS has been developed as the theory of topic-focus articulation (or TFA for short). In FGD's TFA a sentence's linguistic meaning is

partitioned into a contextually given *topic* and a *focus* that is about the topic. The terms topic and focus are based on the structural notion of *contextual boundness*. Each dependent and each head in a sentence's linguistic meaning is characterized as being either contextually bound (CB) or contextually nonbound (NB). Intuitively, items that have been activated in the preceding discourse may function as CB, whereas non-activated items are always considered NB.

contextual boundness

Another ingredient of the FGD framework is the *communicative dynamism*, which defines a (partial) order over the nodes in a sentence's linguistic meaning. The *topic proper* and the *focus proper* are the least respectively most communicatively dynamic elements in a sentence's linguistic meaning. The scale of communicative dynamism has been accounted for the variations in word order (and indirectly tunes) in language like Czech.

In another school of thought, information structure has been studied with the viewpoint of how the message/information is sent or packaged in an utterance. Valduví [1990] defines *information packaging* as "a small set of instructions with which the hearer is instructed by the speaker to retrieve the information carried by the sentence and enter it into her/his knowledge store." Vallduví divides his approaches – to which he refers as 'information articulation', into *topic/comment* approach and *focus/ground* approach. Both (types of) approaches split a sentence's linguistic meaning in two parts.

information packaging

The topic/comment approach splits the meaning into a part that the sentence is about, which is usually realized *sentence-initially*, and a comment. The focus/ground approach splits the sentence's meaning into 'focus' and a 'ground', with the 'focus' being the informative part of the sentence's meaning. The ground anchors the sentence's meaning to what the speaker believes the hearer already knows. The 'focus' expresses what the speaker believes to contributes to the hearer's knowledge. In this sense the 'ground' is also known as 'presupposition' or 'open proposition'.

Further details and discussion on both these schools of thought can be found in Kruijff [2001]. However, it can be observed from these theories that information structure is a presentational means which a speaker employs to indicate to the hearer (*i*) how the sentence's meaning is anchored in the preceding discourse and their mutual beliefs, and (*ii*) how the utterance contributes novel information to the current contex which needs to be added to their mutual knowledge.

Steedman [1996, 2000b,a] develops a theory of information structure in the lines of the Prague School. The particular relevance of Steedman's theory for our work is due to that fact that his theory accounts for both the realization and the interpretation of IS in a sentence's linguistic meaning. The following section provides more insight into his theory of IS.

2.2.1 Steedman's Theme/Rheme

Steedman [1996, 2000b,a] offers a theory of grammar in which *syntax*, *information structure* and *intonational prosody* are integrated into one framework. The un-

derlying claim of this theory is that the surface derivations are associated with a compositional semantics that determines both information structure and predicate argument structure aspects of a sentences's linguistic meaning. Steedman's aim is to provide a information structure-sensitive compositional analysis of English in a Combinatory Category Grammar (CCG). This system is therefore monostratal: the only proper representation of a sentence is the representation of its *linguistic meaning*.

Theme and Rheme

In Steedman [2000a]'s view, the linguistic meaning of an utterance can be divided along two independent information structure dimensions, both of which are relevant to its realization. The first of these dimensions partitions the utterance content into *Theme* and *Rheme*. The theme part links the utterance to the preceding discourse context, and the rheme part advances the discourse by contributing novel information. The bracketing in (20) and (21) (example (4) and (5) from [Steedman, 2000a, pg. 6]) indicates the theme (subscript Th) and rheme (subscript Rh) partitions of the content of the answers in view of the respective questions.

theme
rheme

(20) Q: I know who proved soundness. But who proved COMPLETENESS?
A: (MARCEL)$_{Rh}$ (proved COMPLETENESS)$_{Th}$.
H* L L+H* LH%

(21) Q: I know which result Marcel PREDICTED. But which results did Marcel PROVE?
A: (Marcel PROVED)$_{Th}$ (COMPLETENESS)$_{Rh}$.
L+H* LH% H* LL%

Observe that in (20) and (21) the content of the rheme partition advances the discourse by contributing novel information, whereas the theme links it to the content established by the respective questions. The theme/rheme distinction is similar to the Praguian topic-focus articulation. Informally put, the Steedman's Theme/Rheme partitioning basically tells how the utterance relates to the preceding discourse context.

Steedman formalizes the theme of a sentence as a λ-term involving a functional abstraction. The rheme is a term that can be applied to that abstraction, after which we obtain a proposition. Since CCG is a categorial grammar combining a λ-calculus to represent linguistic meaning, this proposition has the same predicate-argument structure as the composition of the canonical sentence would have resulted in.

For example, the question Q in (21) establishes the theme and can be characterized via functional abstraction using the notation of λ-calculus as follows:

(22) $\lambda x.prove' \; x \; marcel'$

Steedman argues that since the functional abstraction is closely related to the existential operator $\exists$, the notion of theme can be associated with a set of propositions

among those supported by the conversational context that could possibly instantiate the corresponding existentially quantified proposition. Accordingly, for the conversation in (21) the existential in the question Q is the following:

(23) $\exists x.prove'\ x\ marcel'$

The propositions that may instantiate the existential might in a particular context be a set like the following:

(24) $\left\{ \begin{array}{c} prove'\ soundness'\ marcel' \\ prove'\ decidability'\ marcel' \\ prove'\ completeness'\ marcel' \end{array} \right\}$

Steedman refers to such a set as the *rheme alternative set*. The alternative set is, however, not exhaustively known to hearers, and in practice the computation is carried out with quantified expression like (23). Steedman [2000a] presents a model for intonational realization of an utterance's information structure and proposes that it is the choice of the tune employed by the speaker which helps the hearer in establishing the alternative set. The theme tune presupposes the rheme alternative set, whereas the rheme tune restricts the rheme alternative set. The sense in which a theme "presupposes" a rheme alternative set is a pragmatic presupposition, much the same as that in which a definite expression presupposes the existence of its referent.

Following the discussion of [Pierrehumbert and Hirschberg, 1990] (see section 2.1) Steedman identifies the H* L and H* LL% intonation tunes as rheme tune, and L+H* LH% as theme tunes. The intonational contours beneath the respective answer A in (20) and (21) illustrate how these tunes mark the rheme and theme IS partitions. Steedman elaborates on the role of rheme and theme tunes in the discourse context of (21) as follows: The theme tune marked entity PROVE in the answer in (21), establishes the quantified expression in (23) and also presupposes the rheme alternative set in (24). The rheme tune in the respective answer then restricts this rheme alternative set to the proposition $prove'\ completeness'\ marcel'$, which is indicated by the pitch accent marking on COMPLETENESS.

Focus and Background

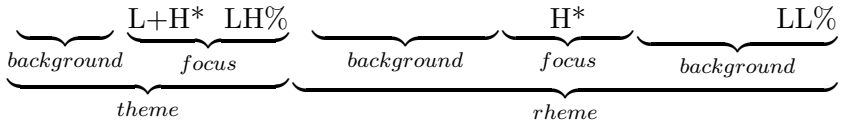
Steedman also defines a second dimension of information structure. This dimension partitions the rheme and the theme IS segments into *focus* and *background* units. Within both theme and rheme segments, those words that contribute to distinguishing the theme and the rheme of an utterance from other alternatives made available by the context may be marked via a pitch accent, while the others are left unmarked. The location of pitch accent(s) in the rheme and theme IS units indicate the words that are in the focus.

Going further, Steedman argues that the theme's focus is optional. There can, but need not, be a marked element in the theme's surface realization. A marked element in the theme is felicitous when the context necessitates contrast with a compatible prior theme.

In this sense, Steedman's partitioning along the second dimension is related to Halliday's Given/New (cited by [Kruijff, 2001]) and to the Praguian division of contextual boundness into contextually bound/contextually nonbound.

The example in (25) (example (14) from [Steedman, 2000a, pg. 11]), below indicates the Theme/Rheme IS partitioning along with their Focus and Background segments. The alignment of the pitch accent H* and L+H* with the contents in the sentence indicate entities, which distinguish the theme and the rheme of the utterance from other alternatives made available by the context.

(25) Q: I know that Marcel likes the man who wrote the musical.
But who does he ADMIRE?
A: (Marcel ADMIRE)$_{Rh}$ (the women who DIRECTED the musical.)$_{Th}$



Steedman argues that the significance of having a pitch accent on *directed* in (25) seems to be in the context offering alternatives that only differ in the relation between *Marcel* and x, as expressed by the quantified expression in the following:

(26) $\exists x.admires'\ x\ marcel'$

The intonational tune in the theme IS unit of the answer in (25) would be infelicitous if the context would not contain an alternative, like the $\exists x.likes'\ x\ marcel'$ we have here. The set of alternative themes provided by the context in (25) is as follows:

(27) $\left\{\begin{array}{c}\exists x.admires'\ x\ marcel' \\ \exists x.likes'\ x\ marcel'\end{array}\right\}$

The kind of alternative set in (27) is what Steedman calls the *theme alternative set*. The theme presupposes this set, and it is the theme tune that restricts it to the existential proposition $\exists x.admires'\ x\ marcel'$. The theme in turn presupposes a rheme alternative set such as the following:

(28) $\left\{\begin{array}{c}admires'\ woman'_1\ marcel' \\ admires'\ man'\ marcel' \\ admires'\ woman'_2\ marcel'\end{array}\right\}$

The rheme of (25), "*the woman who directed the musical*", restricts the rheme alternative set to the proposition $admires'\ woman'_1\ marcel'$. The word DIRECTED is contrasted to distinguish this set from the alternative proposition $admires'\ woman'_2\ marcel'$, which may correspond to say a woman *producer*.

In addition to the partitioning of an utterance's discourse meaning into Theme/Rheme and Focus/Background dimensions, Steedman proposes two further dimensions to information structure (In an unpublished introductory guide on "Using APML to Specify Intonation").

The first of these concerns whether or not the particular Theme or Rheme unit at hand is *mutually agreed* by the speaker and the hearer. Steedman argues that the contentious vs. uncontentious informational status of Theme/Rheme units is realized by the speaker's choice of pitch accent tunes. Following Pierrehumbert and Hirschberg [1990], Steedman identifies the starred L tunes (L*, L*+H, H+L*) as pitch accents indicating the speaker's contentions, whereas starred H tunes (H*, L+H*, H*+L) as indicating the speaker's uncontentiousness.

The second dimension distinguishes the speaker or the hearer as responsible for, "owning", or committed to, the Theme/Rheme IS units. The ownership information status indicates which of the dialogue participants, the speaker or the hearer, has the *ownership* of verifying the truth of the content. Steedman argues that a speaker's claim of commitment to a certain belief may be based upon the actual belief of the agent itself, whereas the speaker's attribution of commitment to the hearer need not be the hearer's actual belief. Steedman suggests that the ownership information status governs the choice of boundary tones that mark the Theme/Rheme IS units. He identifies the L% boundary tones as indicator of a speaker's ownership of an IS unit, and attributes the H% boundary tone to indicate hearer's ownership of an IS unit.

Steedman [1996, 2000b,a] proposes a grammar framework in CCG that allows such information structure-enriched representations of a sentence's linguistic meaning to be realized with surface forms containing intonation contours. In the following section we describe this relationship between intonation and information structure. We will then elaborate the finer aspects of this grammar framework in section 5.1.

2.3 Intonation and Information Structure

In explaining the divergence of phrasal intonation in English from traditional notions of surface syntactic structure Steedman [1991, 2000b] shows that the *intonational structure* in English, essentially as described by Pierrehumbert [1980]; Pierrehumbert and Hirschberg [1990] and others, is directly subsumed by surface syntactic structure, as it is viewed in CCG. The interpretation that the grammar assigns compositionally to the constituents of nonstandard surface derivations directly corresponds to *information structure* of the utterance.

Thus the two possible surface structures for a substring like *Marcel proved completeness* in Figure 2.1 and Figure 2.2 corresponding to (20) and (21) respectively, are due to the differences in their information structure. Steedman claims that the multiple derivations engendered by CCG deliver identical interpretations, which can conveniently be represented as predicate-argument structures or logical forms.

In offering a grammar framework for discourse semantics and intonational prosody, Steedman claims that the information structure encoded in the logical forms has to be inferred from the partial specification implicit in the intonational contours in exactly the same sense that predicate-argument relations have to be inferred from that implicit in the sequence of words. The constituents of the derivation and their

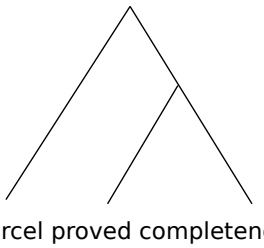


Figure 2.1:
Traditional surface derivation.

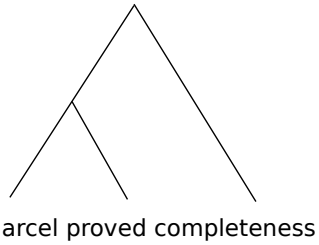


Figure 2.2:
Prosodic surface derivation.

semantic interpretations provide the logical forms that discourse semantic function apply to, and the boundaries of these constituents line up with the intonational structural boundaries.

In the lines of the theory of intonational description of Pierrehumbert [1980]; Pierrehumbert and Beckman [1986] and Pierrehumbert and Hirschberg [1990], Steedman [2000a] defines the intonational contours entirely in terms of two components or "tones": the *pitch accent(s)* and the *boundary*. With this he makes the claim that the *phrasal tunes* in this sense are associated with specific discourse meanings distinguishing information type and/or propositional attitude and that the two independent dimensions of information structure, (see section 2.2.1), are relevant to intonation.

Steedman suggests that in English the Theme/Rheme dimension of information structure contributes among other things to determine the overall shape of the intonational phrasal tune. In particular, the L+H* L H% tune (among others) is associated with the theme, whereas the H* L and H* L L% tunes (among others) are associated with rheme. The second dimension of information structure concerns the distinction between words whose interpretations contribute to distinguish the theme and rheme of the utterance from *alternatives* that the context makes available. This dimension of information structure in English is reflected in the position of the pitch accents themselves. The presence of pitch accents of any type assign salience or contrast independently of the shape of the intonational contour.

In lines of the compositional theory of tune interpretation of Pierrehumbert and Hirschberg [1990], Steedman [2000b,a] also illustrates how the various discourse functions of intonational contours derive from more primitive compositional discourse-semantic elements that are associated with the individual tunes (pitch accents and boundary tones) that make up the contour. By linking the Theme/Rheme, Focus/Background, Agreement and Commitment aspects of a sentence's linguistic meaning with tune meaning, Steedman offers an alternative theory of discourse interpretation as well as intonational realization of information structure meaning. In particular, his main claim is that the theme tunes *presuppose* a rheme alternative set of propositions, and the rheme tunes *restrict* the rheme alternative set to propositions relating to one particular predication which might be instantiated in

a particular context [Steedman, 2000a, pg. 10].

Steedman suggests further that the H* accent yield a rheme not only in combination with LL% boundary tone, but also with a LH% boundary tone. In a similar manner the L+H* pitch accent may occur with LL% as well as LH%. Steedman [2000a] doesn't offer the discourse level distinctions of what Pierrehumbert and Hirschberg [1990] refer to as phrasal accents L and H. However, following the observations in the literature, Steedman claims the the H% boundary mark themes and rhemes alike as the hearer's theme or rheme, whereas L% boundaries mark them as the speaker's.

Steedman's presentation of boundary tones in terms of the ownership of Theme/ Rheme information units is consistent with Pierrehumbert and Hirschberg [1990]'s proposal that a H% boundary requires the hearer to interpret the meaning of an utterance with respect to the utterance that follows, whereas a L% boundary suggest no such "forward reference". Thus the diverse collection of speech acts such as questioning, polite requesting, holding or ceding the turn which have been imputed to H% boundaries in the literature, arise by implication from the marking of information units as hearer's.

2.4 Summary of the chapter

- In this chapter, we provided an overview of the background literature relevant to the realization of intonation in utterances. We started with a review of Pierrehumbert and Hirschberg's compositional theory of tune interpretation [1990]. We described the intonational description followed in their work, and elaborated upon the meaning of the tune elements – *pitch accents*, *phrase accents* and *boundary tones*. Following this, we discussed the compositional meaning of these tune elements.

  The significance of Pierrehumbert and Hirschberg's theory for our work is that their analysis of the contributions of intonational tunes in discourse interpretation offers tune meaning as a means for decoding a speaker's *beliefs*, *intentional* and *attentional state* in a dialogue context. Then, the question that has relevance to the pursuit of research objectives of this thesis is: *Given a representation of discourse semantics, how can we realize it with intonational tunes in the surface forms? And for that matter, how do we encode a speaker's intention, attention and belief state, in such a representation?*

- The theory of *information structure* as proposed by Steedman [2000a] offers a solution to both the issues of *meaning representation* and its *intonational realization*. In the second part of this chapter, we discussed the basic concepts of information structure and deliberated upon Steedman's theory of IS. We described the role of the four dimensions of information structure in presenting the contextually licenced linguistic meaning of an utterance. Next we discussed Steedman's account of intonational realization of an utterance's linguistic meaning.

In Chapter 4, we develop an approach to encode a speaker's intentional and attentional state relative to its beliefs in a sentence's linguistics meaning as its information structure. In Chapter 5, we illustrate our approach to the intonational realization of such a information structure-enriched representation of a sentence's meaning.

3
Theoretical Background

In this chapter, we provide an overview of the software architecture on which we develop and implement our approach to contextually appropriate intonation of clarification requests. We introduce the the architecture schema for the CogX system. We describe the function and process workflow of the system components involved in the interactive continuous learning scenario. Next, we present the logic used to express the semantic representations, called *Hybrid Logic Dependency Semantics*, and discuss its main formal properties for representing information at the various levels of processing. After this, we describe the systemic network approach for utterance content planning. Next, we describe the *Combinatory Categorial Grammar* formalism used for syntactic parsing and realization in our system. Following this we discuss Steedman's approach to combinatory prosody in CCG.

3.1 CogX System Architecture

The CogX system has been developed using the CoSy Architecture Schema (CAS) [Hawes and Wyatt, 2010]. CAS is a set of design principles for developing distributed information-processing software architectures. In this design, the basic processing unit is called a *component*. Components related by their function are grouped into *subarchitectures* (SA). Each subarchitecture is assigned a *working memory*, a message board, which all the components within the subarchitecture may read or/and write to. The inter-component and inter-subarchitecture communication is achieved by writing to and reading from the working memories. The CAS schema is implemented in the CoSy Architecture Schema Toolkit (CAST), which is an open-source, multi-language (Java, C++, Python) implementation of CAS.

CAS

CAST

George is one of the robot in the CogX project on which the system for interactive continuous learning scenarios is being developed and demonstrated (see section 1.3). The aim is to enable George to operate in a real world settings, in communicating with humans and acquiring novel knowledge in a natural way. The demonstrative goal of the George scenario is to show how knowledge can be acquired

during interaction with a tutor in a fully embodied and situated system. The system is composed of three main subarchitectures: the Vision SA, the Binding SA and the Communicative SA (also refered to as Comsys in short).

The **vision subarchitecture** processes visual information, detects and recognizes the objects and makes this information available to other modalities that are part of the cognitive system [Vrečko *et al.*, 2009]. The Vision SA is capable of learning models for visual object properties through dialogue with its human partner. Learning is achieved via two distinct learning mechanism: *explicit learning* and *implicit learning*.

Explicit learning

Explicit learning is a purely tutor-driven learning mechanism that occurs during the initial stage of learning when the robot has no idea of models of world concepts. Explicit learning is therefore used for providing new information or for unlearning of concepts. Implicit learning is triggered by system's own initiative, when it recognizes a gap in its visual knowledge or an opportunity to raise or lower its confidence. Initiating a dialogue with the tutor is a means for the visual subsystem to resolve such knowledge gaps or acquire new information.

Implicit learning

The purpose of the **binding subarchitecture** is to combine modality-specific data representations of the world to a common a-modal one [Jacobsson *et al.*, 2008]. The Binder SA is directly connected to all the subarchitectures in the architecture. It serves as a central hub for the information gathered about entities currently perceived in the scene. Each subarchitecture inserts a *binding proxy* – an a-modal representation of an entity perceived by the subarchitecture – into the binder working memory. The *binding monitor* component in the Binding SA translates the modality-specific representation to an a-modal one. A proxy contains a list of a-modal *binding features*. The binding monitor following a Bayesian network approach determines the correlation of the binding features to connect them together as a *binding union*: a structure that explains that a certain set of proxy refers to the same perceived entity.

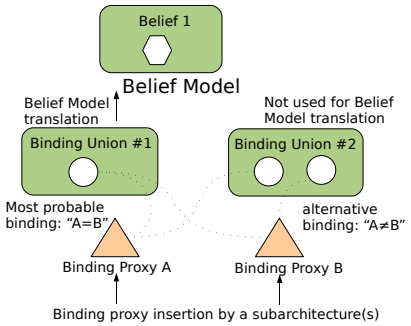


Figure 3.1: Bottom-up construction of beliefs from *Binding Unions*

Given a set of binding unions, the binding subarchitecture constructs and maintains a *belief model*. The model updates occur in a bottom-up fashion on changes in the binding unions, as shown in Figure 3.1.

The **communication subarchitecture** takes care of dialogue processing and

communication with a human. The subarchitecture may be divided into two parts: the *comprehension* and the *production* part. At the **comprehension side**, for speech recognition, we use the Nuance 8.5 speech recognizer. The word lattices provided by the recognizer are then incrementally parsed by a CCG [Steedman, 2000b] parser. The possible parses are then pruned according to the saliency measures on the situated context (visual and dialogue history) provided by the binder so as to reduce the recognition errors [Lison, 2008]. The parse selection in the end yields a most likely parse with a corresponding logical representation, a *logical form*. A logical form is a sorted hybrid logic formula in Hybrid Logic Dependency Semantics (HLDS, see section 3.2) built compositionally during the CCG parsing [Baldridge and Kruijff, 2002].

speech recognition

Comsys has a component called the Continual Collaborative Activity (CCA) (see section 3.3.2), which is situated between the comprehension and the production modules. It takes a pre-processed input from the comprehension components, and produces an output that is picked up and processed by the production side. The CCA tries to assign each recognized utterance an *intention*. To be able to perform such a task, the CCA component uses both the logical representation of an utterance u and the possible situated context references in u to unions (i.e. perceived entities) U on the binder working memory. *Reference resolution* then is performed within the CCA algorithm as part of the abductive explanation of the communicative intention behind u [Kruijff and Janíček, 2009]. Figure 3.2 provides a schema of the comprehension side.

Reference resolution

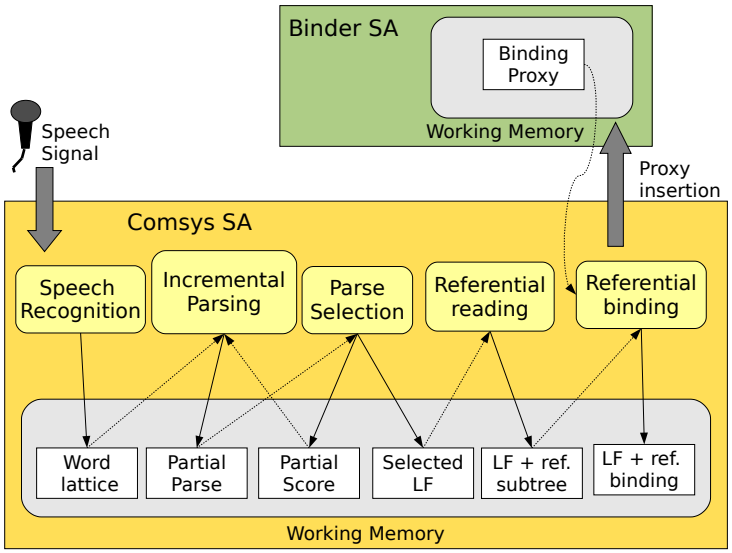


Figure 3.2: Dialogue Comprehension

The process of **speech production** is triggered by the writing of a *proto-logical form* to the Comsys working memory. A proto-logical form is a hybrid logic formula that specifies the *communicative goal* of the utterance to be produced and its content. The proto logical form is processed by the *Utterance Content*

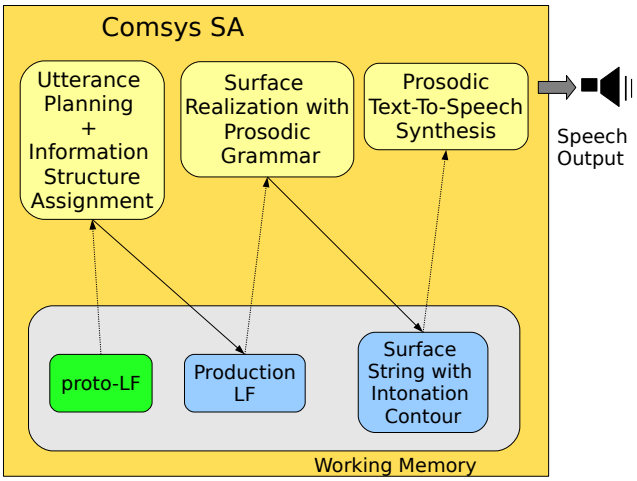


Figure 3.3: Production with Intonation

Planner component and transformed into a logical form that matches the specific linguistic structures and information structure defined in the grammar used for surface realization (see section 3.4).

In Chapter 4 we will discuss in detail our approach to modeling information structure in an utterance's linguistic meaning. Finally the logical form is picked up by the *Surface Realizer*, which is based on statistical realization with CCG [White and Baldridge, 2003; White, 2004]. The realizer uses a grammar that is derived from the grammar used for comprehension extended with the assignment of intonation markers that correspond to the information structure of the utterance. Chapter 5 elaborates on our implementation of the grammar for intonation realization of information structure of an utterance. The resulting surface string containing the intonation marking is then fed to the MARY text-to-speech synthesizer [Schröder and Trouvain, 2003], for speech production. Figure 3.3 provides a schema of the production side in Comsys SA.

3.1.1 Process Workflow

Having introduced the functionality of the core subarchitectures involved in the CogX system for the George scenario, let us take a look at the process workflow among them them. Figure 3.4 provides an overview of the whole system and workflow for an interactive learning scenario. Whenever a new object is introduced to the robot, the Visual SA captures 3D color images and starts processing the images for spaces of interests (SOI). SOIs are then segmented and proto-objects are created. The Object Analyzer component uses these low level proto-objects features to recognize the object properties. The component Visual Mediator then packs all the visual information in a Vision Proxy and sends it to the Binder.

The Binder SA binds the visual information with information from other modal-

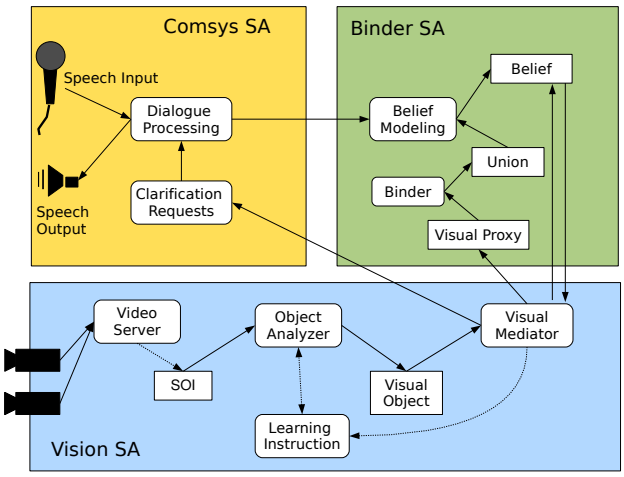


Figure 3.4: Architecture and process workflow

ities and a belief is created from the obtained multi-modal information. When new proxies are introduced, then new visual proxies are posted in the binder and new beliefs are created or some previously acquired beliefs are altered. These beliefs form a basis for further reasoning, learning and communication about the real world entities.

The beliefs acquired in this manner via the visual subsystem can be also altered by the information inputs from the tutor through dialogue via the Comsys SA. The Visual Mediator monitors the beliefs in the Binder SA, waiting for such learning opportunities. When such an opportunity is spotted a learning instruction is sent to the learner, which updates the models (explicit learning). When the visual mediator is uncertain about recognition results, it can send a clarification request to the communication subsystem (implicit learning). The Comsys SA forms a corresponding question and synthesizes it as speech output, and the tutor's answer is then used to update the models.

In the remainder of this chapter we discuss the theoretical background and focus on the key components involved in the production site of the CogX system. We first introduce the HLDS formalism which is used for knowledge representation at various levels (Belief Model, CCA and Dialogue Processing) in our system. Next, we describe the functionality of the Belief Model, the CCA and the Utterance Planner, which are the key components participating in interactive dialogue.

3.2 Hybrid Logic Dependency Semantics

The λ-*calculus* has been used for many years as the standard semantic encoding for categorial grammars and other grammatical frameworks. However, works from Copestake *et al.* [1999]; Kruijff [2001], highlight its inadequacies for both linguistic and computational concerns of representing natural language semantics. One

deficiency of λ-calculus meaning representations is that they usually have to be *type-raised* to the worst case to fully model quantification, and this can increase the complexity of syntactic categories since a verb like *wrote* will need to be able to take arguments with the types of generalized quantifiers.

Kruijff [2001] proposes an alternative approach to represent linguistically realized meaning: namely, as terms of *hybrid modal logic* [Blackburn, 2000] explicitly encoding the dependency relations between heads and dependents, spatio-temporal structure, contextual reference, and information structure. Baldridge and Kruijff [2002] call this unified perspective combining many levels of meaning Hybrid Logic Dependency Semantics (HLDS)

In the following text, which is drawn from Baldridge and Kruijff [2002], we provide a brief discussion on how hybrid logic extends modal logic. Then we describe the representation of linguistic meaning using hybrid logic terms.

3.2.1 Hybrid Logic

Modal logic

Modal logic provides an efficient and logically sound way of talking about *relational structures* as a means of knowledge representation. However it suffers from an inherent "*asymmetry*": the concept of states ("worlds") is at the core of model theory for modal logic, but there is no means to directly *reference* specific states using the object language. This inability of modal logic to indicate where exactly a proposition holds makes modal logic an inadequate representation framework for practical applications like knowledge representation [Areces and Blackburn, 2001] or temporal reasoning [Blackburn, 2000].

Hybrid logic

nominals

Hybrid logic provides an elegant solution to many of these problems by extending standard modal logic while retaining decidability and favorable complexity [Areces *et al.*, 2001; Areces and Blackburn, 2001]. The strategy is to add *nominals*, a new sort of basic formula with which we can explicitly name states in the *object language*. Next to propositions, nominals are therefore first-class citizens in the object language.

Each nominal names a unique state. To get to that state, a new operator, called the *satisfaction operator* is introduced, that enables one to "jump" to the state named by the nominal. Formulas can be formed using both sorts, standard boolean operators, and the *satisfaction operator* "@". A formula, $@_i p$ states that the formula p holds at the state named by i (the nominal). With nominals we obtain the possibility to explicitly refer to the state at which a proposition holds, which is essential for capturing the intuition about temporal reference.

Another interesting characteristic of hybrid logics is the possibility to *sort* the nominals. Sorting enables us to create ontologically rich representations of linguistic meaning. Different sorts of nominals can be introduced in the object language to build up a rich sortal ontology. Additionally, we can reason about sorts because nominals are part and parcel of the object language. This makes it possible to capture the rich ontologies of lexical databases like WordNet in a clear and concise fashion, while retaining decidability and a favorable complexity [Baldridge and

Kruijff, 2002; Areces and ten Cate, 2006].

3.2.2 Representing Linguistic Meaning

Using hybrid logic we can capture three essential aspects of linguistic meaning:

1. *structural complexity*, as modal logic allows us to represent linguistic meaning via sophisticated relational structures
2. *ontological richness*, due to the sorting strategy;
3. the possibility to *refer* to individual states due to the introduction of nominals in the object language.

We can represent an expression's linguistically realized meaning as a conjunction of modalized terms, anchored by the nominal that identifies the head's proposition:

(29) $@_{\mathrm{h:\,sort}}(\mathbf{prop}_h \wedge \langle\delta_i\rangle\,(\mathrm{d}_i : \mathrm{sort}_{d_i} \wedge \mathbf{dep}_i))$

Here the nominal h is the head propositional nominal, which allows us to refer to the *elementary predication* $\mathbf{prop}_h$. The *dependency relations* (such as Agent, Patient, Result, etc.) for the predications are modeled as *modal relations* $\langle\delta_i\rangle$. Each dependent is again labeled by a specific sorted nominal, here d_i with its sort. The features attached to a nominal (e.g. Number, Quantification, etc.) are specified in the same way. Technically, the formula in (29) states that each nominal d_i names the state where a dependent expressed as a proposition $\mathbf{dep}_i$ should be evaluated and is a δ_i successor of h, the nominal identifying the head

The formula in (30) is an example of logical form of the utterance "the box is red". The utterance ascribe the property of being 'red' via the modal relation Cop-Scope (that leads to the nominal red_1) to the subject (refered by nominal box_1) via the modal relation Subject.

(30)
$$
\begin{aligned}
@_{w_1:ascription}(&\mathsf{be} \wedge\\
&\langle\mathsf{Mood}\rangle\ \mathrm{ind} \wedge\\
&\langle\mathsf{Tense}\rangle\ \mathrm{pres} \wedge\\
&\langle\mathsf{Cop\text{-}Restr}\rangle(box_1 : \mathit{thing} \wedge \mathbf{box} \wedge\\
&\qquad\langle\mathsf{Delimitation}\rangle\ \mathrm{unique} \wedge\\
&\qquad\langle\mathsf{Number}\rangle\ \mathrm{sg} \wedge\\
&\qquad\langle\mathsf{Quantification}\rangle\ \mathrm{specific}) \wedge\\
&\langle\mathsf{Cop\text{-}Scope}\rangle(red_1 : \mathit{q\text{-}color} \wedge \mathbf{red}) \wedge\\
&\langle\mathsf{Subject}\rangle(box_1 : \mathit{thing}))
\end{aligned}
$$

Having discussed the HLDS framework for expressing the linguistically realized meaning, we now discuss the functionality of two key components of the CogX system namely Belief Models and CCA, and also see how they employ the HLDS relational structures for representing domain knowledge.

3.3 Multi-Agent Belief Model

belief

A *belief* is the psychological state in which an individual holds a proposition or premise to be true. In the CogX system, Kruijff and Janíček [2009] develop an approach to model agent beliefs. In their model, beliefs reflect an agent's informational state that is related to the agent's understanding of the surroundings. Such an understanding of the environment can be acquired through communication with other agents, as is the case when engaging in information seeking dialogue with the interlocutor, or through direct observations, i.e. as a result of sensory inputs (e.g. from visual subsystem).

From the robot's perspective, beliefs specify what it knows and does not know, about a referent r, e.g., an area in the environment, an object – or, more specifically, relations or properties. *Aboutness* in our system is mainly related to a collection of propositions about a referent's *properties* (color, shape, size, etc.), *category* (type, quality, location, etc.) and relations among them.

Kruijff and Janíček [2009] represented a belief as a formula composed of three parts: *content*, a *spatio-temporal frame* to which it refers, and *epistemic status* indicating the attribution of the belief to other agents.

Belief Content

The content of an agent's belief about a referent r is a proposition, like the following:

(31) referent r is possibly of type box.

(32) referent r has possibly the color red.

(33) size of referent r is not known to me.

A proposition is represented as a logical formula ϕ built up using the HLDS relational structures (see section 3.2). This makes it possible to build up relational structures in which propositions can be assigned ontological *sorts*, and referred to by using *indices* (the nominal variables). These HLDS relational structures are then employed in formulating agent beliefs.

Following this, the propositions in (31) and (32) are represented as logical formulas (LFs) in (34) and (35) respectively.

(34) $@_{\text{e1: entity}}(\mathsf{r}_{e1} \wedge \langle \textit{Property} \rangle\, (\text{t1} : \textit{type} \wedge \mathsf{box}))$

(35) $@_{\text{e1: entity}}(\mathsf{r}_{e1} \wedge \langle \textit{Property} \rangle\, (\text{c1} : \textit{color} \wedge \mathsf{red}))$

The LF in (34) is a representation of a robot's belief that the referent r is an *entity*, and that referent r is possibly of type box, as indicated by the relational feature *Property*. On the other hand, the LF in (35) represents the robot's belief that the referent r is an *entity*, and this referent r has possibly the color red.

One obvious advantage of using HLDS relational structures for representation is that instead of requiring to maintain multiple semantically unrelated propositions in a system, as the following in (36) a property-type *index* variable, such as c1 in (35), could simply take domain values in the range of its ontological sort *color*.

(36) a. referent r has possibly the color red.

b. referent r has possibly the color yellow.

Epistemic Status

The epistemic status of a belief refers to the attribution of the belief to other agents. It basically tells how beliefs have been acquired. Origin-wise beliefs can have the following three epistemic status classes:

- ***private beliefs*** are those that arise from within a system. For a robot, this corresponds to an interpretation of sensory input (e.g. from visual subsystem) or deliberation.
- ***attributed beliefs*** are interpretations of another agent's actions or utterances. They can be acquired e.g. when the human tutor provides some information about an object, or performs a physical action. The robot, however, is cautious so as not to internalize the attributed beliefs immediately. The only way the robot is able to internalize a belief attributed to another agent is through *grounding*.
- ***shared beliefs*** are those attributed beliefs which become part of a *common ground* between two agents following the process of grounding.

Beliefs thus specify not only what the robot knows about a referent r, but also what the robot presumes to be shared knowledge about r, and what the robot presumes other agents could know about r.

The epistemic status of a belief is represented as part of the belief formula. A private belief of an agent a_1 about propositional content ϕ (as in LF (34)) is expressed as $(K\,\{a_1\}\,\phi)$. K is an operator denoting the relation between ϕ and a_1, and $\{a_1\}$ is a non-empty set of agents. On the other hand, a shared or mutual belief, held among many agents, is expressed as $(K\,\{a_1, a_2, a_3, ...a_n\}\,\phi)$. A belief content ϕ attributed to an agent a_1, but not yet mutually agreed upon by agent b_1 is expressed as $(K\,\{b_1[a_1]\}\,\phi)$.

To account for a belief's situatedness, Kruijff and Janíček [2009] relate beliefs to *spatio-temporal frames*, which are essentially intervals in space and time. The main purpose of spatio-temporal frames in their current approach is to establish the interdependence of spatio-temporal frames. That is, if a belief holds in a spatio-temporal frame σ, then it should also hold in all the spatio-temporal frames $\sigma' \subseteq \sigma$.

Formally speaking, beliefs are formulas that assign content formulas epistemic statuses and spatio-temporal frames. The formula in (37) is a representation of a belief comprising these three aspects, and is read as, "agent a_1 holds a belief ϕ during the spatio-temporal frame σ."

(37) $K\,\{a_1\}\,/\sigma : \phi$

3.3.1 Uncertainty in Beliefs

The interesting aspect of modeling beliefs as HLDS relational structure is that nominals can be defined as multi-valued state variables [Brenner and Kruijff-Korbayová, 2008]. The absence of a value for an index state variable (a nominal) is interpreted as *ignorance*, not as falsehood. For example, the logical form in (38) with property-type state variable t1 having no domain value implies that the agent does not know the *type* property of the referent r, not that the referent has no *type* (as per a closed-world assumption of classical planning where: what is not known to be true is *false*).

(38) $@_{\text{e1: entity}}(\mathsf{r}_{e1} \wedge \langle \mathit{Property} \rangle\, (\text{t1} : \mathit{type}))$

This way HLDS relational structures allow us to represent gaps in an agent's beliefs. Furthermore, the state variables can be quantified over, for example using the ? to express a *question* about the state variable. The LF in (39) represents a quantified belief that questions the *type* property of the referent r.

(39) $?t1.@_{\text{e1: entity}}(\mathsf{r}_{e1} \wedge \langle \mathit{Property} \rangle\, (\text{t1} : \mathit{type}))$

The presence of such a quantified belief in an agent's belief model may lead to an intention to seek the *type* value of the referent entity r from the human interlocutor.

uncertainty

The approach to use quantified logical formulas further allows us to represent an agent's *uncertainty* in acquired beliefs. This is achieved by using the ? operator to quantify such a belief. For example, the quantified LF in (40) represents an agent's uncertainty that the referent r is of *type* box:

(40) $?t1.@_{\text{e1: entity}}(\mathsf{r}_{e1} \wedge \langle \mathit{Property} \rangle\, (\text{t1} : \mathit{type} \wedge \mathsf{box}))$

The quantified LF in (40) can be extended further to also represent an agent's uncertainty due to the presence of multiple hypotheses about a referent or piece of knowledge. For example, it can happen that the perceptory sensors provide the robot with multiple hypotheses about a referent's color property. A hypothesis that a referent r has possibly the color red or possibly the color orange can then be formulated as a belief in (41).

(41) $?c1.@_{\text{e1: entity}}(\mathsf{r}_{e1} \wedge \langle \mathit{Property} \rangle\, (\text{c1} : \mathit{color}\,(\langle \mathit{List} \rangle \mathsf{disjunction} \wedge \langle \mathit{First} \rangle (\text{c2} : \mathit{color} \wedge \mathsf{red}) \wedge \langle \mathit{Next} \rangle (\text{c3} : \mathit{color} \wedge \mathsf{orange}))))$

The LF in (41) follows White *et al.* [2004b]'s notion of $\langle \mathit{List} \rangle$ to represent alternative values for a property-type variable c1, which can take domain values in the range of its ontological sort *color*.

3.3.2 Continual Collaborative Activity

Since sensory perception is necessarily subjective, partial and therefore uncertain, a robot may believe that it sees something, and can usually say how sure it is about this referent. However, it might not be sure whether other agents (the tutor for

that matter) perceive the same environment the same way. Grounding through dialogue is a means to mitigate this uncertainty. Kruijff and Janíček [2009] model grounding in agent beliefs as a *Continual Collaborative Activity* (CCA), in which beliefs - private, attributed and shared - are continually subjected to *verification*. It is the presence of gaps, uncertainties or ambiguities in the robot's belief model that give rise to communicative intentions to seek clarification or information for their resolution. The possibility to initiate a dialogue with other agents during the verification process is a means for an agent to clarify any uncertainty or fill knowledge.

verification

An important aspect of the verification process in CCA is the notion of *assertions*, which Kruijff and Janíček [2009] employ to relax the restriction of Stone and Thomason's [2002] "Principle of Coordination Maintenance", which (roughly) says "what the speaker says is what the hearer understands". This is a very strong assumption – in fact, it says that no grounding is required. In order to remove this, Kruijff and Janíček [2009] use the mechanism for belief revision, and assertions. These mechanism enable the robot's to use them as the *backchannel feedback* responses to give the human user a reason to believe that the robot believes the given fact.

assertions

The CCA component is therefore responsible not only for grounding in dialogue but also for initiating a request for communication – for clarifications, informations questions, acknowledgement and assertions. Figure 3.5 provides an overview of the workflow for the realization of a communicative intention raised as part of the CCA in an agent's belief model. In the following section we provide an overview of the

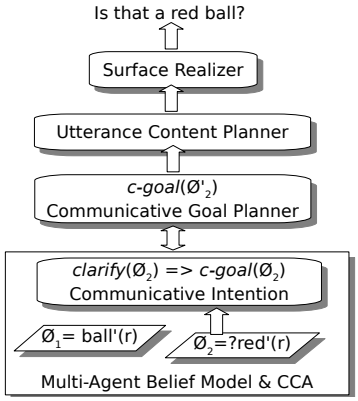


Figure 3.5: Realizing a Communicating Intention

functionality of the *Utterance Content Planner* component, which is responsible for planning the predicate-argument structure and semantic representation for a given *communicative goal*. In Chapter 4 we will elaborate the functionality of the *Communicative Goal Planner*, and also revisit the content planner for encoding the information structure semantics as part of an utterance's meaning representation. In Chapter 5 we elaborate on the *Surface Realizer* component.

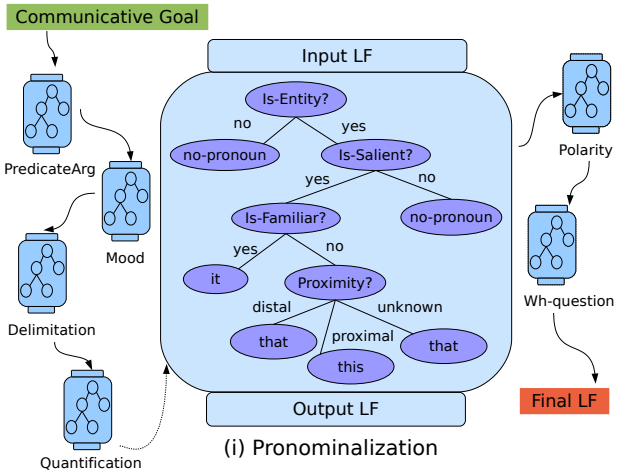


Figure 3.6: Utterance Planning with Systemic Network

3.4 Utterance Content Planning

planning grammar

The utterance content planner component in the CogX system is based upon Kruijff's [2005] approach to generating contextually appropriate utterances. The planner uses a *planning grammar*, which is a systemic network in the tradition of generation systems for systemic functional grammar [Mathiessen, 1983; Bateman, 1997]. These systems take an abstract logical form as input (e.g. a LF representing a communicative goal), and enrich it with specifications of the desired linguistic realization. In a systemic network:

- A *system* represents a paradigmatic choice, i.e. a choice about an aspect of the meaning to be specified by the logical form we are planning. A system consists of an *entry condition*, *actions* associated with the different choices the associated *chooser* can make, and an *output*.
- Both the *entry condition* and the *output* of a system take the shape of a HLDS logical form, and an indication of the *locus* within that logical form. The combination of a system's output logical form and the corresponding locus may form the entry condition for another system. It is in this way that we obtain a network of systems. Figure 3.6 is a fragment of the planning grammar used in our system.
- We specify the decision process involved in a system as a *decision tree* or *chooser*, associated with the system. In the chooser, we can pose several inquiries about the logical form and the contextual status of discourse referents, to guide the decision process.
- On the basis of the choice we make, the system performs one or more *operations* on the logical form, to expand it; we thus reflect grammatical features

directly as content in the logical form.

The system (i) in Figure 3.6 provides an example of a system, namely pronominalization. The point of the system is to specify a pronomial referential expression for the *entity* at the current locus. The decision tree indicates the inquiries, which the chooser associated with this system can make. With queries like Is-Entity, Is-Salient, Is-Familiar, the chooser successively gathers information from various models in the system as to whether or not the entity is *salient* in the current context, and whether or not it has been introduced earlier. With the query Proximity the choosers gather the *spatial* information as to whether the entity being referred to is far or close.

The leaf nodes in the tree indicates the actions the system may take on the LF at hand. One of the possible actions of the decision tree is it, indicating that the *entity* at current locus should be referred to with *it*. This results in the system performing the following actions on the logical form:

1. add-proposition: adds the proposition *it* to the nominal of the current locus.
2. add-feature: adds a new feature *Number* with value sg (for singular).

Besides these two operations, there are other operations that a system can specify depending on the sort of current locus. The operation add-relation adds a new sorted relation between the locus and a some nominal. Operation assign-type specifies the sort of the locus. The operation identify-nomvar, identifies a nominal for a specified named relation of the locus, and then moves the locus to this nominal. We can define for a variable whether it is to have system-local scope, or global scope. This way, we can reference other parts of a logical form, outside the scope of the subtree that is currently in the locus.

These operations define the basic inventory for extending a logical form through *substitution*(cf. [Kruijff, 2005, pg. 4]).

1. add-feature := $@_{\text{n:nomv}}\ \phi \Rightarrow\ @_{\text{n:nomv}}\phi\,\&\,@_{\text{n:nomv}}\ \langle Feat\rangle\,(\text{value})$
2. add-proposition := $@_{\text{n:nomv}}\ \phi \Rightarrow\ @_{\text{n:nomv}}\phi\,\&\,\text{prop}$
3. add-relation := $@_{\text{n:nomv}}\ \phi \Rightarrow\ @_{\text{n:nomv}}\phi\,\&\,@_{\text{n:nomv}}\ \langle Rel\rangle\, n' : nomv'$
4. assign-type := $@_{\text{n:nomv}}\ \phi \Rightarrow\ @_{\text{n:type}}\phi$

The main way the utterance planner brings in *context-sensitivity* is through the types of inquiries posed in a chooser. In the architecture, the utterance planner runs as an agent with access to the various short-term and long-term models that the robot maintains for the environment it is situated. Each of these models is equipped with attentional mechanisms, which model current attentional prominence or salience (like models of the discursive or visual context).

Based on the results of these inquiries, we can decide how to reflect the contextual status of this entity in the logical form. An alternative way of making

contextual information available to the chooser is by specifying the contextual information as features-values in the LF itself. We opt for this alternative when the modules raising a communicative request have a direct access to relevant contextual information. This makes the information readily available for the chooser, thereby reducing the overhead on the choosers to fetch the specific information.

Contextual appropriateness of an utterance may take various forms, not just by appealing to the appropriate (and maximally informative) modal context to refer, but also formulating the information structure to reflect attentional status, and utterance coherence in the current discourse. In Chapter 4 we detail our approach to modeling information structure in the linguistic meaning of the utterance.

The end result of utterance content planner is a logical formula (Final LF, see Figure 3.6), which is processed by the Surface Realizer component for realizing the linguistic meaning with surface forms. The surface realizer in our system is based upon the CCG framework [Steedman, 2000b]. In the following section we provide a brief overview of the CCG framework for grammar development, and motivate why it is suitable for modeling prosodic derivations.

3.5 Combinatory Categorial Grammar

Combinatory Categorial Grammar (CCG) was first introduced by Ades and Steedman [1982] as a generalization of the earlier Categorial Grammar (CG) frameworks of Adjukiewicz [1935] and Bar-Hillel [1953]. It was further expanded by Steedman [1996, 2000b]. Like other forms of categorial grammars, CCG share a strong commitment to the *principle of compositionality*, [Frege, 1892; Montague, 1974] – that is, the assumption that the meaning of a complex expression is determined by its *structure* and the *meanings of its constituents*. Stated otherwise, syntax and semantics are *homomorphically related* and may be derived in tandem.

lexicon

CCG is a *fully lexicalized* grammar formalisms. That is, all the language-specific information is contained in the *lexicon*, not in the grammar rules, as in classical context-free grammars. The responsibility of defining the syntactic forms lies at the level of individual lexical items. This allows the syntactic behavior of each morpheme (in terms of subcategorization frame, possible features and constraints, multi-word combinations, etc.) to be specified with great precision, leading to a fine-grained grammatical description.

CCG combines the advantages of being both *linguistically very expressive* and *efficiently parsable*. It has been used to model a wide range of linguistic phenomena [Steedman, 2000b; Steedman and Baldridge, 2009], most notably a unified – and very elegant – treatment of *extraction* and *coordination*. CCG has a completely *transparent interface* between surface syntax and the underlying semantic representation, including predicate-argument structure, quantification and information structure.

Lexicon

As in other varieties of Categorial Grammar, elements like verbs are associated with a syntactic category that identifies them as functions and specifies the type and directionality of their arguments and the type of their result. A "*result leftmost*" notation is used, in which a rightward-combining functor over a *domain* β into a *range* α is written as α/β, and the corresponding leftward-combining functor is written $\alpha\backslash\beta$. α and β may themselves be function categories. The lexical categories can be augmented with an explicit identification of their semantic interpretation via a colon operator.

For example, a transitive verb is a function from (object) NPs into predicates, which are themselves functions from (subject) NPs into S:

(42) $\mathsf{proved} := (\mathsf{S}\backslash\mathsf{NP})/\mathsf{NP} : \mathit{prove}'$

Combinatory Rules

Functor categories such as (3.5) can combine with their arguments using the *combinatory rules*. The most basic combinatory rules are the forward and backward applications.

(43) *Forward Application* (>) := $(\mathsf{X}/\mathsf{Y}) : f \quad \mathrm{Y} : a \quad \Rightarrow \mathrm{X} : fa$

(44) *Backward Application* (<) := $(\mathsf{X}\backslash\mathsf{Y}) : f \quad \mathrm{Y} : a \quad \Rightarrow \mathrm{X} : fa$

The semantic interpretation of this and all other combinatory rules is completely determined by its syntactic form under the following principle [Steedman, 2000b]:

(45) *The Principle of Type Transparency*
All syntactic categories reflect the semantic type of the associated logical form, and all syntactic combinatory rules are type-transparent versions of one of a small number of simple semantic operations over functions including application, composition and type-raising.

The derivation in (46) illustrates the application of these combinatory rules and also the semantic composition.

(46)

Marcel	proved	completeness	
$NP : marcel'$	$(S\backslash NP)/NP : prove'$	$NP : completeness'$	
		$(S\backslash NP) : prove'\,completeness'$	>
$S : marcel'\,prove'\,completeness'$			<

The derivation yields an S with a compositional interpretation, equivalent as usual under the convention of left associativity to $(prove'\ completeness')\ \ marcel'$.

Rather than invoking rules of deletion or movement, CCG allows certain further operations on functions which are called functional composition rules.

(47) *Forward Composition* (>B) := $\mathsf{X}/\mathsf{Y} : f \quad \mathsf{Y}/\mathsf{Z} : g \Rightarrow \mathrm{X} : \lambda x.f(gx)$

(48) *Backward Composition* (<B) := $\mathsf{X}\backslash\mathsf{Y} : f \quad \mathsf{Y}\backslash\mathsf{Z} : g \Rightarrow \mathrm{X} : \lambda x.f(gx)$

Combinatory grammars also include type-raising rules, which turn arguments into functions over functions-over-such-arguments. These rules allow arguments including subjects to compose, and thereby take part in coordination, as in *Marcel proved, and I disproved, completeness.*

(49) *Forward Type-Raising* (>T) := $\mathsf{X} : a \Rightarrow \mathsf{T}/(\mathsf{T}\backslash\mathsf{X}) : \lambda f.fa$

(50) *Backward Type-Raising* (<T) := $\mathsf{X} : a \Rightarrow \mathsf{T}\backslash(\mathsf{T}/\mathsf{X}) : \lambda f.fa$

Here, X ranges over argument categories such as NP and PP. T is a metavariable that schematizes over a number of instantiations subject to a restriction that the functions T/X and T\X must be categories consistent with the parameters of the language in question. For example, English NPs can raise over S\NP, (S\NP)/ NP, S/NP, and so on, but not (S\NP) \ NP. The rules have an "order-preserving" property. For example, (49) turns an NP into a rightward looking function over leftward functions and therefore preserves the linear order of subjects and predicates. The interpretation of such rules is again entirely determined by the Principle of Type Transparency (45).

The significance of this theory for present purposes follows from Steedman's [2000a] argument that, if in order to account for coordination and relativization we take strings like *you think that Marcel proved* to be surface constituents of type S/NP, then they must also be possible constituents of non-coordinate sentences like *Marcel proved completeness*, which must permit derivation (51), as well as the traditional derivation (46), which with type-raising appears as in (52):

(51)
$$\dfrac{\dfrac{\dfrac{\dfrac{\text{Marcel}}{NP : marcel'}}{S/S\backslash NP}\,{>}\mathsf{T}}{: \lambda f.f\ marcel'} \qquad \dfrac{\dfrac{\text{proved}}{(S\backslash NP)/NP : prove'} \qquad \dfrac{\dfrac{\text{completeness}}{NP : completeness'}}{(S\backslash(S/NP)}\,{<}\mathsf{T}}{: \lambda p.p\ completeness'}}{(S\backslash NP) : \lambda x.prove'x\ marcel'}\,{>}\mathsf{B}}{S : marcel'prove'completeness'}\,{<}$$

(52)
$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\text{Marcel}}{NP : marcel'}}{S/S\backslash NP}\,{>}\mathsf{T}}{: \lambda f.f\ marcel'} \qquad \dfrac{\dfrac{\text{proved}}{(S\backslash NP)/NP : prove'} \qquad \dfrac{\dfrac{\text{completeness}}{NP : completeness'}}{(S\backslash NP)\backslash((S\backslash/NP)/NP)}\,{<}\mathsf{T}}{: \lambda p.p\ completeness'}}{(S\backslash NP) : \lambda y.prove'y\ completeness'}\,{<}}{S : marcel'prove'completeness'}\,{<}$$

It is important to notice that once we simplify or "normalize" the interpretations by β-reducing λ-abstracts with arguments, as we have tacitly done throughout these and earlier derivations, both yield the same appropriate proposition $prove'\ completeness'\ marcel'$.

Steedman [2000a] argues that the relevance of these nonstandard surface structures that were originally introduced to explain coordination in English is simply that they subsume the intonation structures that are needed in order to explain the possible intonation contours for sentences in (53) and (54).

(53) Q: I know who proved soundness. But who proved COMPLETENESS?
A: (MARCEL)$_{Rh}$ (proved COMPLETENESS.)$_{Th}$
H* L L+H* LH%

(54) Q: I know which result Marcel PREDICTED. But which results did Marcel PROVE?
A: (Marcel PROVED)$_{Th}$ (COMPLETENESS.)$_{Rh}$
L+H* LH% H* LL%

Intonational boundaries, when present as in spoken utterances A in (53) and (54), contribute to determining which of the possible combinatory derivations such as (51) and (52) is intended. The interpretations of the constituents that arise from these derivations, far from being "spurious," are related to semantic distinctions of information structure and discourse focus. In Chapter 5, we discuss Steedman's model of combinatory prosody which accounts for intonational realisation of information structure in a CCG framework. We also elaborate our implementation of this model in the OPENCCG platform.

3.6 Summary of the chapter

- In this chapter, we provided an overview of the CogX system architecture. We have introduced the core subarchitectures participating in the interactive continuous learning scenario, namely the *Visual SA*, the *Binding SA* and the *Communicative SA*. Next we elaborated on the key dialogue processing components of the Communicative SA. These involve the Belief Model component, the CCA and the Utterance Planner. We discussed Kruijff and Janíček's [2009] approach to Multi-Agent Belief Modeling, and Continual Collaborative Activity for *grounding* in dialogue. Then we discussed Kruijff's [2005] approach to generating context-sensitive utterance for realizing an agent's communicative intention. We also described the Hybrid Logic Dependency Semantics, which is the logic we use in the Communicate SA for *semantic representations*.

- In Chapter 4, we discuss our approach to producing contextually appropriate intonation in an agent's utterances. We will illustrate how an agent's intentional and attentional state can be employed in assigning the *information structure* status to the agent beliefs underlying a communicative intention. In order to encode the information structure semantics we will extend the planning grammar introduced in Utterance Planner (section 3.4) with the *systems* for information structure assignment.

- In the last section of this chapter we provided an overview of the Combinatory Category Grammar (CCG) and discussed the main *combinatory rules* and illustrated some non-traditional surface derivations. These derivations, as Steedman argues, enables us to model the orthogonal surface derivations of prosodic phrases. In Chapter 5, we model such derivations in OpenCCG platform to achieve the *combinatory prosody* proposed by [Steedman, 2000a].

Part II

Approach

4

Modeling Information Structure

In this chapter, we describe our approach to producing contextually appropriate robot utterances. We start by laying out the conceptual framework for modeling *information structure* using a robot's *belief state*. Next, we describe our approach for deriving the informational states of a robot's beliefs, and for encoding their discourse meaning as information structure. Following this, we elaborate on the implementation details of our approach. We end with an illustration of how our approach links together the surface structure, semantic and information structure in one single representation.

4.1 Contextual Appropriateness

The very purpose of a communicative act is to convey information so that it is readily and clearly understood. Acts of communication, however, do not take place in a vacuum, but rather in the context of a larger discourse. It is therefore important for the meaning of an utterance to be coherent with the preceding and the current context. The notion of *information structure* (IS) in an utterance's linguistic meaning helps us explain the utterance's coherence in this larger discourse.

information structure

In Chapter 2, we have presented the basic concepts of IS and its formulation in various theories. In this chapter we use this notion of information structure for addressing the issue of contextually appropriate realization of a robot's utterances. Towards this, we follow Steedman's [2000a] theory of IS, and develop an approach for deriving the necessary aspects of the IS partitioning from a robot's belief state.

The reasons we follow Steedman's formulation of IS are: (*i*) the IS partitioning of an utterance content into the *Theme/Rheme* allows us to represent how the utterance relates to the preceding and the current dialogue context, (*ii*) the partitioning of the Theme and the Rheme parts further into *Focus/Background* segments allows us to distinguish the most salient aspects of the current dialogue context from the others, (*iii*) the *Agreement* dimension allows us to specify whether or not the dialogue participants mutually agree to believe in the particular Theme or Rheme unit at hand, and (*iv*) by specifying the information status along the *Ownership* dimension, we are able to distinguish the speaker or the hearer as responsible for

"owning", or being committed to, these information units.

Example (55) (adopted from [Steedman, 2000a, pg. 27, (68)]) illustrates the linguistic signs (in CCG), that participate in the derivation of the sentence *Marcel proved completeness*.

(55)

Marcel PROVED		COMPLETENESS
	L+H* L H%	H* L L%
$(\mathsf{S}_\phi \backslash \mathsf{NP}_\phi)$		$\mathsf{S}_\phi \backslash (\mathsf{S}_\phi / \mathsf{NP}_\phi)$
$: [H^+](\theta(\lambda x.{}^*prove' x\ marcel'))$		$: [S^+](\rho(\lambda p.p\ {}^*completeness'))$

Here the θ (for Theme) and the ρ (for Rheme) marking in the semantic dimension of the signs respectively indicate the Theme and Rheme partitions of the utterance content. The Focus and Background aspects of the Theme/Rheme partitions is respectively indicated by the presence and absence of the '*' (asterisk) mark on the corresponding propositions. The $[H]$ and $[S]$ modalities respectively indicate the commitment of the hearer and the speaker to the Theme/Rheme content. The superscript '+' (for agreed) and '-' (for not-agreed) on these two modalities indicate whether the hearer and the speaker mutually agree to believe in these propositions. In this manner, the four dimensions of the information structure contribute in presenting the contextually appropriate meaning of this sentence in a dialogue context.

We believe that the contextual appropriateness of a robot's utterances can also be established by accounting for these four aspects of information structure in an utterance's linguistic meaning. The question that needs to be addressed first is: *Given a model of the dialogue context where do we derive these four aspects of informativity in a robot's utterance from?*

In Chapter 3 (see section 3.3), we discussed that in an interactive continuous learning setup, a robot's utterance is a realization of a communicative intention to convey, seek or clarify some piece of information about some referent. The intention to communicate arises as a consequence of an intended update of the robot's *belief state*, set against the background of already available/formed beliefs, and relative to some referent(s). Therefore, the contextual appropriateness of a robot's utterances has to inevitably also reflect the robot's *intentional* and *attentional state*, relative to the updates in its belief state.

belief state

In view of these four dimensions of informativity and the informational state of a robot's beliefs, we propose that the contextual appropriateness of a robot's utterance can be established by accounting for the following four questions during the information structure assignment:

(56)
1. How does an utterance exhibit the informational state of the underlying beliefs with respect to the *common ground* established among the dialogue participants?
2. How does an utterance present the informational state of the underlying beliefs with respect to robot's *attentional state*?
3. How does an utterance reflect upon a robot's attitude like *contentions*

about the underlying beliefs, which may or may not have been yet *grounded*?

4. How does an utterance indicate a robot's claim of *commitment* to the propositional content of the underlying beliefs?

Otherwise stated, how does the meaning conveyed by an utterance account for the robot's belief state in the context of current dialogue. Let us take each of these questions in turn and discuss the aspects of a robot's belief state that an utterance needs to be able to present, how the contextual appropriateness of an utterance relates to a robot's belief state, and how the informational state of the robot's beliefs can be represented in Steedman's formulation of information structure.

4.1.1 Agent Beliefs and Common Ground

In an interactive learning setup, based on whether or not the robot believes a particular belief to be mutually known among the dialogue participants, it can be classified as a *shared* or *unshared* belief. Moreover, a robot's belief that some piece of information is also known to the dialogue participants could be based on the fact that they have actually arrived at such a mutual understanding by *grounding* it to the *common ground*, or it could be simply a pragmatic presupposition on the robot's part about the participants' knowledge.

common ground

The contextual appropriateness of an utterance therefore has to account for distinguishing between beliefs, which the robot believes (or assumes) to be the common ground among the dialogue participants and those which it believes (or assumes) to be not yet grounded.

From the perspective of the dialogue participants, the robot's expression of these distinctions allows them to access its cognitive state. That is, the belief state of the robot, the intentions behind the utterance, and what the robot believes or assumes to be their mutual knowledge.

Now, how does an utterance present the distinctions between the part of the utterance meaning that has been already established and the part which is novel in the current dialogue context? Following Steedman [2000a]'s Theme/Rheme IS partitioning, we propose that:

Definition 1. *The beliefs underlying a communicative intention which the robot believes (or assumes) to be mutually known and hence established in the dialogue context, should be assigned the discourse information status of Theme, whereas those which it intends to be grounded among the dialogue participants should be assigned the discourse information status of Rheme.*

The *Theme/Rheme* information structure distinction in an utterance's linguistic meaning is realized by a set of different intonation tunes associated with them. In this manner the informational state of the robot's beliefs with respect to the common ground can be incorporated in the decision process for assigning a contextually appropriate intonation tune for an utterance.

4.1.2 Agent Beliefs and Attentional State

In an interactive learning setup, a robot and a human tutor usually switch their discussion from one referent to another, from one property to another property of a referent, and even over properties and relations across various referents. Throughout such interactions the robot's beliefs associated with these referents and their properties change their *attentional state*. The robot's beliefs attain *saliency* in the dialogue context when the referents or properties corresponding to them are *to be made* or are already the most *salient* aspects of the ongoing interaction.

saliency

The contextual appropriateness of an utterance should therefore also account for distinguishing between referents which the robot intends to be made the most salient in the current dialogue context and those which is assumes to be already salient. The ability of a robot to state which of the referents and what aspects of these referents among the others are the focus of the current dialogue helps in reducing the scope for any situational ambiguity for the dialogue participants.

Now, how does a robot's utterance draw the attention of the dialogue participants to the intended referents? Following Steedman [2000a]'s Focus/Background IS partitioning, we propose:

Definition 2. *The beliefs underlying an utterance that correspond to the referents, which the robot intends to be made salient among the other alternative referents in the current dialogue context, should be assigned the discourse information status of Focus. The beliefs corresponding to the referents, which the robot believes to be already salient or unambiguously established in the current context, should be assigned the discourse information status of Background.*

The *Focus/Background* information structure distinction in an utterance's linguistic meaning is reflected in the position of the pitch accents in the intonational contour. The presence of a pitch accent of any type is generally agreed to assign salience or contrast independently of the shape of the intonational tunes. In this manner we make provision for the attentional state of a robot's beliefs in contributing to the decision process of assigning a contextually appropriate intonation to an utterance.

4.1.3 Agent Beliefs and Uncertainty

In an interactive learning setup a robot acquires information about its surroundings through various modalities (such as vision, dialogue) or from deliberations. However, sensory perception (visual or audio) is necessarily subjective, partial and therefore *uncertain*. Initiating a dialogue with the human partner to seek a clarification is a means for a robot to resolve such uncertainties. To achieve this, the robot should be able to indicate what aspects of a referent or referent properties it believes to be doubtful.

The contextual appropriateness of a robot's utterance has to therefore also account for presenting a robot's *uncertainty* about the propositional content of a

uncertainty

belief. From the perspective of the human tutor, the expression of uncertainty by the robot is an indication that the robot's perception of the world is different from the actual reality. This may prompt the tutor to verifying his own perception of the world, or take actions towards clarifying the robot's uncertainty.

So how does a robot's utterance present its uncertainty in understanding certain aspects of its surroundings? Following [Steedman, 2000a]'s notion of *mutual agreement* over IS units, we propose:

Definition 3. *The beliefs underlying a communicative intention for which the robot has partial or no certain knowledge should be assigned the discourse information status of Uncertain, while the rest should be assigned the discourse information status of Certain. These informational aspect of a belief proposition should be specified along the Agreement dimension of the Theme and Rheme IS units.*

The informational status of an utterance content as to whether or not the speaker is certain about it, is realized by different set of pitch accent tunes in the intonation contour. In this manner, by specifying the robot's uncertainty about a belief proposition as part of the utterance's linguistic meaning, we provide for it to contribute to the decision process of determining a contextually appropriate intonation for an utterance.

4.1.4 Agent Beliefs and Claim of Commitment

An robot's beliefs or knowledge can be divided into two subdomains, namely: a set S of propositions that the robot (the speaker) claims to be committed to, and a set H of propositions which the robot claims the hearer(s) to be committed to. By commitment, we mean to say if the speaker claims to be *in a position* to know whether the proposition expressed is *true*. The informational state of an agent's belief with respect to the commitment state indicates whether the speaker takes the *ownership* of such a claim, or attributes the claim to the hearer. ownership

A speaker's claim of commitment to a certain propositional content may be based upon the actual belief of the agent itself, however, the speaker's attribution of commitment to the hearer need not be the hearer's actual belief. The contextual appropriateness of an utterance have to therefore also account for distinguishing between beliefs, which the robot claims to be commitment to and those to which the robot attributes the claim on the hearer(s).

The ability of a robot to express these distinctions allows it to convey to the dialogue participants that, who in the current context has the responsibility of what. By committing itself to a belief the robot indicates its owning to the responsibility of verifying the truth of the expressed proposition, whereas by attributing the responsibility to the hearer the robot indicates its lack of commitment to the proposition expressed.

So how does a robot's utterance present who in the dialogue context has the authority of what? Following [Steedman, 2000a]'s notion of *ownership*, we propose that:

Definition 4. *The beliefs underlying a communicative intention to which the robot claims to be committed to should be assigned the discourse information status of speaker owned, whereas those for which the robot attributes the claim to the hearer should be assigned the discourse information status of hearer owned. These informational aspects can be specified along the Ownership dimension of the Theme and Rheme IS units.*

It is the intonation of an utterance that establishes the discourse meaning, which specifies which of the participants have the ownership of the utterance content. Intonation contours with a *rising boundary tone* (as in information questions and clarifications) suggest the speaker's lack of commitment to the content, and therefore indicate the speaker's attribution of the authority on to the hearer, whereas an intonation contours with a *falling boundary tone* (as in responses to information questions and assertions) indicate the speaker's commitment to the information content.

The **Definition 1** to **Definition 4** outline, at a very basic level, the conceptual framework for presenting the various contextual informational aspects of a robot's beliefs. However, in realizing the contextual appropriateness of an utterance these four dimensions of informativity do not act in isolation, but rather interact with each other and with the robot's intentional and belief state. For example, a robot's *uncertainty* in a belief proposition, and hence the lack of commitment to it, invariably also requires that the *ownership* of the proposition be attributed to the the hearer. In such a scenario we observe that the IS assignment rule in **Definition 4** needs to be mindful of the rule in **Definition 3**.

In the following section we revisit these basic definitions and discuss possible accounts of such interplay of a robot's intentional and belief state in an interactive learning scenario. Our aim here is to lay down the general principles based on which we make the concrete decisions for information structure assignment in robot utterances. In section 4.3, we elaborate our implementation for presenting contextual appropriateness of robot utterances in the George scenario.

4.2 Assigning Information Structure

4.2.1 Theme/Rheme Information Status

In **Definition 1**, we proposed that beliefs, which a robot believes (or assumes) to be the mutual knowledge of the dialogue particpants should be assigned the discourse information status of Theme, whereas those which the robot intends to be grounded to the mutual knowledge should be assigned the discourse information status of Rheme. However, the presentation of some part of an utterance's meaning as shared (the theme) and some as *to be shared* (the rheme) is rather a decision choice, which a speaker makes in a given context.

For example, the rheme of an utterance need not always be novel in the dialogue context. A belief that is mutually established in preceding dialogue context may

be presented again as a rheme in some other context. Typical of such a scenario is when a piece of mutually established information is presented as a response to a *wh-question*, or when the robot intends to seek further information or clarification about some already established information. In such a scenario, the assignment of *Theme/Rheme* information structure status to a robot's beliefs requires the following alternatives to **Definition 1**:

IS Rule 1. *A mutually known belief when being presented as an assertion or as an answer to a previously asked question should be assigned the discourse information status of Rheme.*

IS Rule 2. *A mutually known belief when being presented as a subject about which the speaker seeks further information or clarification, should be assigned the discourse information status of Rheme.*

Note that an agent's belief that a piece of information is also known to the hearer need not always be based upon the fact that they have mutually established it as a common ground, but can also be a *pragmatic presupposition* that is consistent with the context. In such a scenario the speaker presumes the hearer to accommodate the yet unshared information as a mutual belief. Accordingly, the Theme/Rheme IS assignment requires the following alternative rule to **Definition 1**:

presupposition

IS Rule 3. *A hitherto unshared belief when being presented as mutual knowledge, thereby suggesting presupposition on part of the speaker that the hearer makes an accommodation, should be assigned the discourse information status of Theme.*

4.2.2 Focus/Background IS Status

As discussed in **Definition 2**, the purpose of assigning a belief the discourse information structure status of *Focus* is basically to distinguish or *contrast* the corresponding referents from other alternative referents that are salient in the current dialogue context.

The use of contrast is intrinsic and is an integral part of the presentation when a piece of information is presented for grounding in the current dialogue. However, for a belief that the robot assumes or presupposes to be mutually established, use of contrast is mandated only when some other mutually established belief(s), due to its prior existence or accommodability in the context, can also be an alternative instantiation for the presentation. Therefore, we need an alternative to the *Focus/Background* assignment rule in **Definition 2** for handling the assignment of contrast in Theme information units.

IS Rule 4. *If a referent or referent property is unambiguously established in the current context, it should be assigned the discourse information status of Background. However, if there exists alternative (or accommodable) mutually established referents of the same category, it should be assigned the discourse information status of Focus in order to distinguish it from these alternatives.*

With the Focus/Background assignment rules in **IS Rule 4** and **Definition 2** we are able to distinguish the Rheme and the Theme of an utterance from the set of other alternatives in the current context. However, it is important to note that when multiple referents of a category are present (or accommodable) in the context, contrast should be assigned to those propositions about the already salient (or to be made salient) referent which *uniquely* distinguish it from the others.

For example, if there exists a red and a blue ball in the current context. Then in order to make the red ball the most salient aspect in the current context, the discourse information status of Focus should be assigned to the propositional content 'red', as it is this property which distinguishes it from the other (blue) ball. The property 'ball' being unambiguously established should be assigned the status of Background. The following rule specifies the Focus/Background assignment in such a scenario.

IS Rule 5. *If there exist other salient referents or properties of the category of the referent, which the speakers intends to be made most salient, then only those propositions of this referent which uniquely distinguish it from other alternatives, should be assigned the discourse information status of Focus. The remaining propositions should be assigned the status of Background.*

This rule also applies to the referents which the speaker believes to be mutually established and already salient in the context.

IS Rule 6. *If there exist other salient referents or properties of the category of the referent, which the speakers believes to be already salient or unambiguously established, then only those propositions of this referent which uniquely distinguish it from other alternatives, should be assigned the discourse information status of Focus. The remaining propositions should be assigned the status of Background.*

4.2.3 Agreement State and Informativity Status

In **Definition 3**, we proposed that the uncertainty of a robot in a belief proposition due to partial or complete lack of knowledge can be represented as an informational aspect of the presentation using the *Agreement* dimension of the IS units. On the other hand, it is also possible that during the interactive learning the robot might find some discrepancy in the information conveyed to it by the tutor. This may be due to a perceived difference between the robot's *private beliefs* and the *attributed beliefs*. In such an eventuality, if the robot intends to express its disagreement to ground the proposition expressed by the tutor, then the contextual appropriateness of the utterance should also account for pointing out the particular piece of the information that it believes to be *contentious*.

disagreement

contentious

We believe that a robot's perceived (or apparent) disagreement over a belief proposition can also be realized in an utterance by specifying the informational state against the Agreement dimension of the Theme/Rheme IS units as follows:

IS Rule 7. *A speaker's disagreement over a belief proposition should be specified with the discourse information status of Disagreed against the Agreement dimension of the respective Theme/Rheme unit. On the other hand, the belief propositions about which the speaker doesn't have any contentions, should be assigned the discourse information status of Agreed.*

The information structure status as to whether the speaker agrees or disagrees to a propositional content is realized with different set of pitch accent tunes in the intonational contours for the respective information unit.

4.2.4 Ownership Informativity Status

We proposed that the *Ownership* information status of a robot's belief can be derived from its informational state as to whether the robot is in a position to know the truth of the propositional content. While a robot's claim of commitment to a belief may be motivated from its actual beliefs or knowledge, its attribution of the responsibility to the hearer may be simply a presupposition, or based on the contributions of the hearer to the common ground. Following these, the claim of commitment to a belief proposition as per **Definition 4** can be more specifically stated as:

IS Rule 8. *A belief proposition which originated in the speaker's belief model should be assigned the Ownership information structure status of speaker owned.*

IS Rule 9. *A belief proposition which originated in the hearer's belief model should be assigned the Ownership information structure status of hearer owned.*

Arguably, a speaker's claim of commitment to a particular belief proposition need not be only due to its own actual knowledge, but could also be based on attributed beliefs that have been grounded and established as mutual knowledge. For example, a robot's response to a human query (possibly for verification) could be based upon some previously established beliefs. In such a scenario the speaker may choose to present the knowledge as its own.

IS Rule 10. *A mutually known belief when being presented as an assertion or as an answer to a previously asked question may be assigned the Ownership information structure status of speaker owned.*

Continuing with the current argument, a speaker's attribution of the ownership of a belief proposition to the hearer need not always be based on the fact that the belief originated with the hearer, but it may also be the speaker's presupposition that the hearer believes in the expressed proposition. For example, a speaker may commit the hearer to own the claim of some belief proposition, which originated with the speaker, but due to partial or complete lack of knowledge the speaker is not sure of committing himself. The attribution of the ownership to the hearer may be an attempt on the speaker's part to elicit a clarification from the hearer on the propositional content. This is because, if the speaker's presupposition that

the hearer believes in the particular proposition is wrong then the hearer may offer a counter-evidence to the speaker claim. However, if the hearer doesn't offer any such counter claim, the speaker may as well assume the propositional content as their mutual belief, for the time being.

Typical for such a scenario are clarification requests, where the speaker attributes the ownership of the expressed proposition to the hearer to elicit a response. The following rule accounts for the IS assignment in such an eventuality:

IS Rule 11. *A belief proposition about which the speaker has partial or no certain knowledge should be assigned the Ownership information structure status of hearer owned.*

This is also applicable to scenarios when the speaker intends to express disagreement over belief propositions.

IS Rule 12. *A belief proposition which the speaker disagrees to accept should be assigned the Ownership information structure status of hearer owned.*

The speaker's presupposition that a unshared information is mutually believed should also be presented as hearer owned.

IS Rule 13. *A hitherto unshared belief when being presented as mutual knowledge, thereby suggesting presupposition on part of the speaker that the hearer makes an accommodation, should be assigned the Ownership information structure status of hearer owned.*

In the CogX system, the implementation for producing contextually appropriate robot utterances is based on these general principles. In the following section we describe the details of this implementation.

4.3 The Implementation

Figure 4.1 provides a scheme of the overall process for producing contextually appropriate robot utterances in our system. Compare this process workflow with the one in Figure 3.5. In addition to the other components, we now see an interplay of the four aspects of informativity in utterance content planning. In section 3.3, we have introduced the functionality of the Multi-Agent Belief Model and the CCA component. In section 3.4, we have discussed the approach to context-sensitive utterance content planning in the CogX system. In the following sections we discuss the task of planning a communicative goal, and then revisit the task of utterance content planning, albeit for presenting the contextual appropriateness based on the information structure, as motivated in the previous section.

4.3.1 Communicative Goal Planning

In a interactive learning setup, a robot's utterance is a reflection of its intention to *convey*, *seek* or *clarify* some piece of information about some referent. Now,

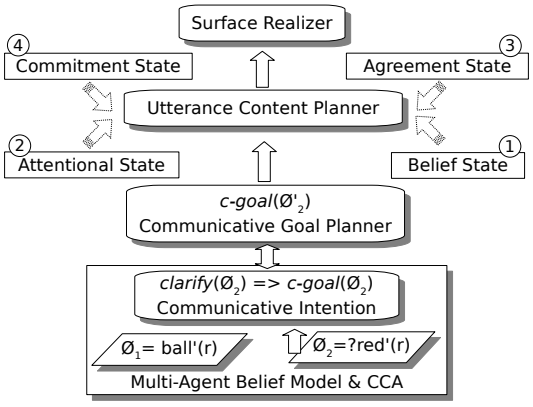


Figure 4.1: Architecture schema for realising a Communicative Intentions

an utterance that realizes a robot's (the speaker) communicative intention needs to be able to at least indicate (*i*) the *referent* the speaker is referring to, (*ii*) the *aspects* of this referent that the speaker is interested in, and (*iii*) the speaker's intention(s) as to what it intends to achieve with the utterance. These information details are essential for planning an utterance's content because they correlate with the *predicate-argument relations* of the sentence being planned. For example, the referent of a robot's communicative intention corresponds to the *subject* of a sentence, whereas the aspect or property of this referent that the speaker intends to ascribe to this referent correlates with the *predicate*. The communication intention indicate whether the predicate-argument relation is to be presented as a question or a clarification, or as an assertive response.

Representation

The relationship between a communicative intention, the referent and the property the robot intends to ascribe to this referent are represented as an HLDS structure. As described in section 3.2.2, this is achieved by first introducing a nominal, which acts as a discourse vantage point (dvp) for the communicative event in the ongoing discourse. Next, the dependency relations between the predication and the arguments is modeled as *modal relations*. We refer to such an elementary relational structure (or proto-logical formula) as a *communicative goal*. The proto-logical formula in (57) illustrates an example communicative goal.

communicative goal

$$
\begin{aligned}
(57)\quad @_{d:dvp}(\mathsf{c\text{-}goal} \;&\wedge \langle SpeechAct\rangle \mathsf{clarification} \\
&\wedge \langle Content\rangle (a1 : ascription \\
&\qquad \wedge \langle Target\rangle (e1 : entity \wedge \mathsf{r}) \\
&\qquad \wedge \langle Property\rangle (c1 : color \wedge \mathsf{red})))
\end{aligned}
$$

The nominal d is the head propositional nominal of sort dvp, which allows us to refer to the elementary predication c-goal. The speaker's intention underlying

the predication is modeled via the dependency relation *SpeechAct*, while the actual content of the predication is modeled via the dependency relation *Content*. The predicate-argument relations of the sentence in turn are also modeled as dependency relations *Target* and *Property*, which respectively correlate with the referent and the property the speaker intends to ascribe to this referent. These dependency relations and their sortal values in the communicative goal in (57) represent the robot's intention to *clarify* whether *ascribing* the color *red* to some referent r is alright.

Communicative Goal Planning

A communicative goal can be planned by any component in the CogX system that is in a position to use dialogue for information exchange. Depending on the information at its disposal, a component may be able to specify all the relevant information for a communicative goal. If the component doesn't have the information at its disposal, it may either inquire the relevant models in the system for the details, or formulate a proto-logical formula with whatever information it has access to, and request the *Communicative Goal Planner* (CGP) component to do the remaining part of the goal planning. For example, observing the communicative goal in (57), we notice that the component seeking dialogue (possibly) didn't have much information about the referent r, with which it could be referred to (e.g. its *type* or *shape*).

The CGP component in the CogX system is responsible for planning a particular communicative goal. The CGP fetches the relevant information both from the feature-values already present in the proto-logical formula (that came in as a request) and from inquiries to the various models in the system (cf. section 3.4).

Communicative Intention

communicative intention

The CGP models the communicative intention via the dependency relation *SpeechAct*. Depending on its functionality, a component in the CogX system may have recourse to initiate dialogue for *clarification*, *seeking information*, *asserting* beliefs, *acknowledgement* and *greeting*. When planning a communicative goal, a component maps the respective dialogue action choice via the dependency relation *SpeechAct*.

For example, the CCA component in a robot's belief model is able to infer from a quantification operator ? in the belief formula that the propositional content is *contentious*, as in (58a) (cf. section 3.3.1). Requesting the tutor to *clarify* the contentious proposition, as in (58b) is one of the possible actions for the CCA to resolve it. This triggers the CCA component for the communicative goal planning (Figure 4.1 illustrates this workflow), wherein the communicative intention for clarification is modeled via the dependency relation *SpeechAct*, as illustrated in (58c).

(58) a. $\phi_2 := ?c1.@_{\text{e1: entity}}(\mathsf{r}_{e1} \wedge \langle \mathit{Property} \rangle\, (\text{c1} : \mathit{color} \wedge \mathsf{red}))$

b. $\mathsf{clarify}(\phi_2)$

c. $@_{d:dvp}(\mathsf{c\text{-}goal} \wedge \langle SpeechAct\rangle \mathsf{clarification})$

Next, the component (or the CGP) needs to model the content of the sentence via the modal relation *Content*. A component may use the explicit knowledge about the referent and the predicate, or exploit the *sort* information of the nominals corresponding to the discourse referents. The nominal types such as *thing*, *entity*, *type*, *physical*, etc. in the sortal ontology suggest that the particular nominal refers to some physical entity in the environment which may be identified as the subject of a sentence. On the other hand sort values such as *color*, *shape*, *size*, etc. suggest that a nominal refers to some quality of a physical entity which may be correlated with the predicate of a sentence. The *referent* and its *predicate* can then be modeled via the dependency relations *Target* and *Property* respectively under the modal relation *Content*.

referent
predicate

For example, with the sortal values of nominals in the quantified belief proposition in (58a), the CCA component is able to infer that the nominal c1 (of sort *color*) is some quality which is to be ascribed to the referent r, which is identified by the nominal e1 (of sort entity). Having identified these nominal, the CCA models them as the dependency relations *Target* and *Property* respectively and updates the communicative goal (58c), which results into the proto-logical form in (57).

In this way, enabling a component to specify the relevant details (at its disposal) in the proto-logical form reduces the burden of system inquiries on the CGP while allowing us to centralize the task of goal planning. As an illustration of CGP's role, the proto-logical form in (57), may be processed further by the CGP to gather sufficient information about the referent r in order to be able to refer to it unambiguously. For example, CGP may request the multi-agent belief model for additional propositional content related to the referent nominal e1. A belief proposition such as (59) contained with the agent's belief model is one such piece of additional information about nominal e1 that the CGP can utilize for specifying the referent.

(59) $\phi_1 := @_{\mathrm{e1:\,entity}}(\mathsf{r}_{e1} \wedge \langle Property\rangle\,(\mathrm{t1} : type \wedge \mathsf{ball}))$

Using the sort value *type* of the nominal t1, the CGP infers that this property of the discourse referent e1 can be also used to refer the referent r. The CGP incorporates this change by transforming the modal relation *Target* as follows:

(60) $@_{d:dvp}(\mathsf{c\text{-}goal} \wedge \langle SpeechAct\rangle \mathsf{clarification} \wedge \langle Content\rangle(a1 : ascription \wedge \langle Target\rangle(e1 : entity \wedge \mathsf{ball}) \wedge \langle Property\rangle(c1 : color \wedge \mathsf{red})))$

In this manner CGP transforms the communicative goal in (57) into a informationally richer structure. With these dependency relations and their sortal values in (60), the communicative goal represents the robot's intention to *clarify* whether *ascribing* the color *red* to the referent *ball* is alright.

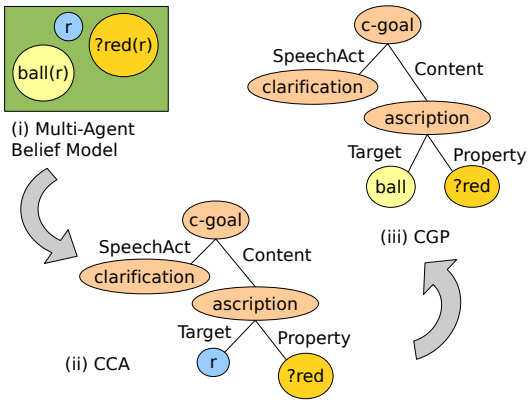


Figure 4.2: Communicative Goal as a relational structure.

Figure 4.2 illustrates the scheme for formulating a communicative goal as a HLDS relational structure. Observe how the propositional content of a robot's beliefs (maintained in Multi-Agent Belief Model) are augmented as the predicate-argument relations in a relational structure by the CCA and the CGP. As discussed in section 4.1.1 and 4.2 the informational status of these belief propositions have a crucial role in presenting the contextually appropriate meaning of the utterance. In the following section we discuss how these informational aspects are encoded for a sentence's verbal head and its arguments.

4.3.2 Encoding Information Structure in Linguistic Meaning

systems

In this section we extend the planning grammar, introduced in section 3.4, with *systems* that encode the information structures semantics in the linguistic meaning of an utterance.

Encoding Theme/Rheme Information Structure Status

We start with the assignment of the Theme/Rheme IS status to a verbal head in a proto-logical from. Following this we will describe the systems for assigning the Theme/Rheme IS status to the subject and predicate arguments of the verbal head.

Event type nominals

Following **Definition 1**, the Theme/Rheme information status of a discourse event (represented by the verbal head) can be derived from its informational state with respect to the common ground. That is, whether the robot believes or assumes the particular discourse event to be novel in the current dialogue or mutually established. The following pairs of dialogue exemplifies the role of the informational state of a verbal head.

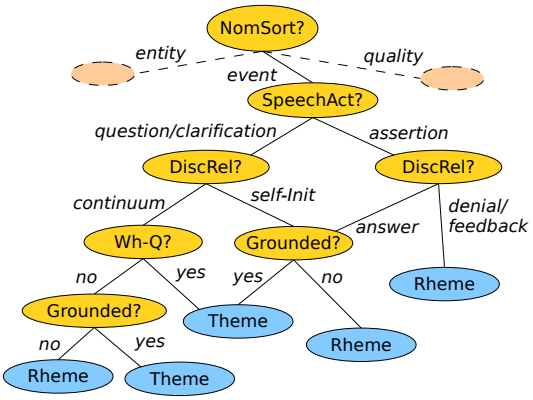


Figure 4.3: A *system* for assigning Theme/Rheme IS status to *event* type nominals.

(61) a. H: (*places a red ball*)
R: (What COLOR is the ball?)$_{Rh}$

b. H: (*places a red ball*)
R: (Is the ball RED?)$_{Rh}$

c. H: (*places a red ball*)
R: (This is a BALL)$_{Rh}$

d. H: (*places a red ball*)
R: (I see a RED thing.)$_{Rh}$

e. H: (*places a red ball*) This is a ball.
R: (Is)$_{Rh}$ the ball)$_{Th}$ (RED?)$_{Rh}$

f. H: (*places a red ball*) What color is the ball?
R: (The ball is)$_{Th}$ (RED)$_{Rh}$

g. H: (*places a red ball and a green box*) Which object is the red?
R: (The BALL)$_{Rh}$ (is RED)$_{Th}$

One thing that is common in these examples and many other similar scenarios is that the verbal head in a clarification or information question is part of the *Rheme* information unit. This is due to the fact that a clarification request introduces a new event in the dialogue context. In (61a)-(61e) the robot's clarification requests suggest that the robot believes that the color of the referent object is not a common ground among the dialogue participants. Hence it is presented as the Rheme of the utterance. The same applies for robot's back-channel feedback or assertive responses (61c) and (61d).

On the other hand the verbal head in responses to human query, as in (61f) and (61g), usually form part of the *Theme* information unit. This is due to the fact that the preceding wh-question or dialogue establishes the discourse event as a common ground.

The decision tree in Figure 4.3 illustrates the *system* responsible for assigning the Theme/Rheme information structure status to a nominal representing the verbal

head of the sentence under planning. The decision tree models the following IS assignment rules:

IS Rule 14. *The verbal head of a self-initiated information question or clarification request should be assigned the information status of Theme if the robot believes it to be grounded, otherwise Rheme.*

IS Rule 15. *The verbal head of a wh-question about a referent, last mentioned, should be assigned the information status of Theme.*

IS Rule 16. *The verbal head of a clarification request about a referent, last mentioned, should be assigned the information status of Theme if the robot believes it to be grounded, otherwise Rheme.*

IS Rule 17. *The verbal head of a assertive response should be assigned the information status of Theme if the robot believes it to be grounded, otherwise Rheme.*

IS Rule 18. *The verbal head of a denial response should be assigned the information status of Rheme.*

IS Rule 19. *The verbal head of a feedback response should be assigned the information status of Rheme.*

A system such as the one in Figure 4.3 is provided access to the contextual informational aspects through the features that are available in the communicative goal itself, or via inquiries to various models in the architecture. For example, the response to a query like *SpeechAct*? (represented by a tree node in the decision tree) can be fetched from the dependency relation feature *SpeechAct* that is already present in the proto-logical form. On the other hand, the query *DiscRel*? basically inquires a model in the system's architecture regarding the discourse relation of the communicative utterance at hand with the preceding discourse. That is, whether the communicative intention is to answer a previously asked question or seek further information or clarification about a previously mentioned referent, or it is simply a feedback response.

The query *Grounded*? inquires the informational state of a referent with respect to the common ground. In our system, this information is derived from the *epistemic status* of agent's beliefs corresponding to that referent. The epistemic status indicates whether a particular belief is *shared* or *private* or *attributed*. As discussed earlier in section 3.3, the epistemic class status of a belief is represented as a formula, such as those in (62).

(62) For a belief proposition ϕ_1 at hand, a formula such as:

a. $K\,\{robot, tutor\}\,\phi_1$ indicates that it is *shared* by robot and the tutor.

b. $K\,\{robot\}\,\phi_1$ indicates that it is the robot's *private belief*.

c. $K\,\{robot\,[tutor]\}\,\phi_1$ indicates that the robot *attributes* it to the tutor.

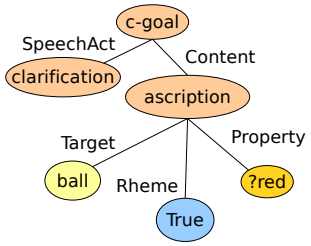


Figure 4.4: Extended proto-logical from with Theme/Rheme informativity status.

Thus a belief with epistemic class status *shared* indicates that the robot assumes the propositional content to be grounded, whereas propositional content with *private* or *attributed* class status to be not-grounded.

The Theme/Rheme informational status of the referent is then specified by extending the proto-logical form with an informational feature. The leaf nodes in a decision tree, (cf. Figure 4.3), indicate the feature values to be added to the proto-logical form via the add-feature operation (see section 3.4). The Rheme IS status is specified by feature-value pair *Rheme*=true, whereas the Theme IS status is specified with feature-value pair *Rheme*=false.

As an illustration, the proto-logical form planned by the CGP in Figure 4.2 is updated by the action choice following the rule **IS Rule 14**. The relational structure in Figure 4.4 with the Theme/Rheme feature-value pair *Rheme*=true illustrates the extended proto-logical from.

Entity type nominals

The decision tree in Figure 4.5 and Figure 4.6 illustrate the *systems* that assign the Theme/Rheme information structure status to the arguments of a verbal head, which have the nominal sort of a physical entity. The referent 'ball' in the robot utterances in (61a) and (61e) is an example of a nominal of this sort. Following the **Definition 1** these systems assign the Theme/Rheme informational status as follows:

IS Rule 20. *An entity type subject argument of the verbal head in a wh-type question/clarification should be assigned the information status of the Rheme. On the other hand, an entity type object should be assigned the status of the Theme if the robot believes it to be grounded, otherwise the Rheme.*

IS Rule 21. *An entity type argument of the verbal head in a non wh-type question/clarification should be assigned the information status of the Theme if the robot believes it to be grounded, otherwise the Rheme.*

IS Rule 22. *An entity type argument of the verbal head in an assertive response should be assigned the information status of the Rheme if it matches the sort of entity sought in the previously asked question.*

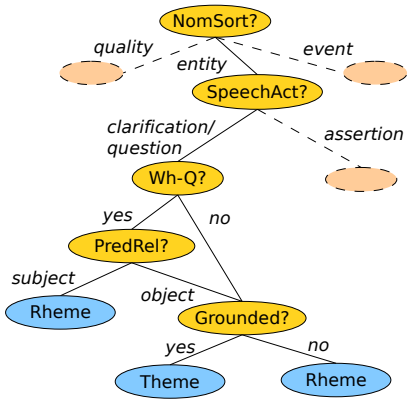


Figure 4.5: A *system* for assigning Theme/Rheme IS status to *entity* type nominals.

IS Rule 23. *An entity type argument of the verbal head in an denial response (negative response) should be assigned the information status of the Theme if the robot believes it to be grounded, otherwise the Rheme.*

IS Rule 24. *An entity type subject argument of the verbal head in a feedback response should be assigned the information status of the Rheme. On the other hand, an entity type object should be assigned the status of the Theme if the robot believes it to be grounded, otherwise the Rheme.*

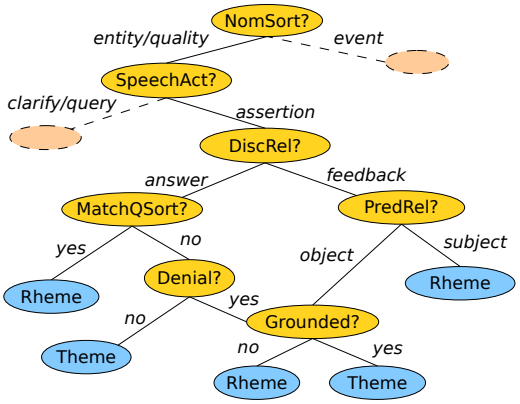


Figure 4.6: A *system* for assigning Theme/Rheme IS status to *entity* and *quality* type nominals.

Quality type nominals

The decision tree in Figure 4.6 and Figure 4.7 illustrate the *systems* that assign the Theme/Rheme information structure status to the arguments of a verbal head which have the nominal sort as *quality* or *property*. The referent property 'color' and

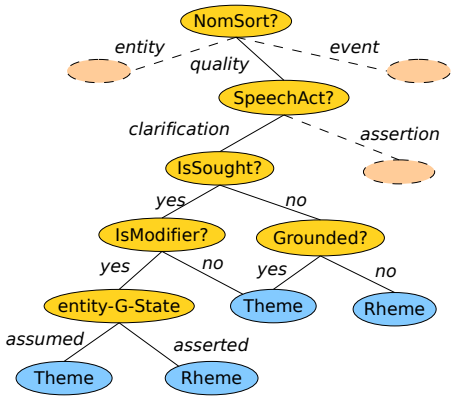


Figure 4.7: A *system* for assigning Theme/Rheme IS status to *quality* type nominals.

quality 'red' in the robot utterances in (61a) and (61e) are examples of nominals of these sorts. Following the **Definition 1**, these systems assign the Theme/Rheme informational status as follows:

IS Rule 25. *A quality type argument of the verbal head in an assertive response should be assigned the information status of the Rheme if it matches the sort of entity sought in the previously asked question.*

IS Rule 26. *A quality type argument of the verbal head in an denial response (negative response) should be assigned the information status of the Theme if the robot believes it to be grounded, otherwise the Rheme.*

IS Rule 27. *A quality type subject argument of the verbal head in a feedback response should be assigned the information status of the Rheme. On the other hand, a quality type object should be assigned the status of the Theme if the robot believes it to be grounded, otherwise the Rheme.*

IS Rule 28. *A property type subject argument of the verbal head in a wh-type question/clarification should be assigned the information status of the Rheme. On the other hand, a property type object argument should be assigned the status of the Theme if the robot believes it to be grounded, otherwise the Rheme.*

IS Rule 29. *A property type argument of the verbal head in a non wh-type question/clarification should be assigned the information status of the Theme if the robot believes it to be grounded, otherwise the Rheme.*

For example, the proto-logical form in Figure 4.4 is extended by the systems in Figure 4.5 and Figure 4.7 for specifying the Theme/Rheme information status for the subject and the predicate (the quality type nominal) in the proto-logical from. The extended relational structure for the proto-logical form is illustrated in Figure 4.8.

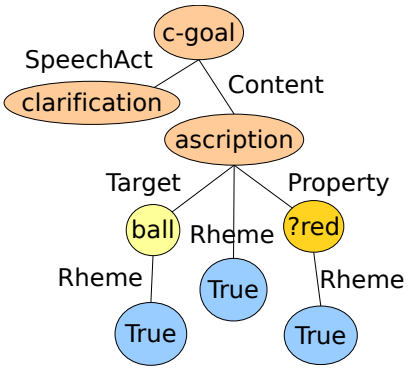


Figure 4.8: Extended proto-logical from with Theme/Rheme informativity status.

Encoding Focus/Background Information Structure Status

The systems assigning the Focus/Background information status take as an input the proto-logical form extended by the systems for assigning Theme/Rheme IS status. This allows us to employ the IS Theme/Rheme feature *Rheme*. This provides for exploiting the decisions of the inquiries made by the earlier systems, for deriving the Focus/Background informational aspects. In this thesis the scope of a robot's clarification does not cover clarifications requests about the discouse events. The robot is mainly concerned with the object types and its properties. Therefore, we simply resort to assigning a verbal head the default information status of the Background.

The decision tree in Figure 4.9 and Figure 4.10 illustrate the *systems* responsible for assigning the Focus/Background information structure status to entity and quality type nominal arguments of a verbal head of the sentence under planning. Following the general principle in **Definition 2**, the systems model the following specific IS rules for our implementation:

IS Rule 30. *A non pre-modified noun type argument of the verbal head bearing the discourse information status of Rheme should also be assigned the information status of Focus.*

IS Rule 31. *A non pre-modified noun type argument of the verbal head bearing the discourse information status of Theme should be assigned the discourse information of Focus if the referent object is to be made salient, otherwise Background.*

IS Rule 32. *A pre-modified noun type argument of the verbal head should be assigned the discourse information status of Focus if the referent object is to be made salient, otherwise Background.*

While the **IS Rule 30** provides for the rheme-focus marking for the utterances such as (63a), **IS Rule 32** allows for the *unmarked* accent placement in (63b), and **IS Rule 31** provides for contrastive theme-focus marking, as shown in (63c).

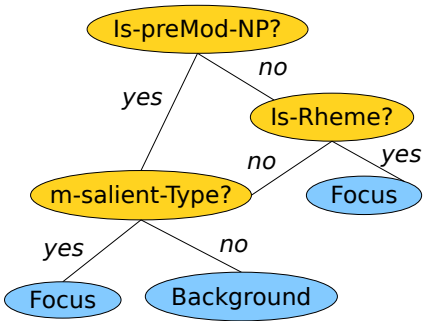


Figure 4.9: A system to assign Focus/Background to *entity* type nominal.

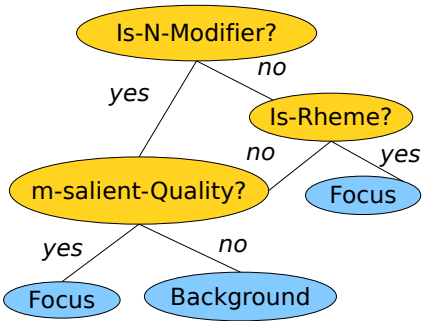


Figure 4.10: A system to assign Focus/Background to *quality* type nominal.

(63) a. R: (Is that a BALL)$_{Rh}$
b. R: (Is that a red BALL)$_{Rh}$
c. R: (The BALL is)$_{Th}$ (RED?)$_{Rh}$

The system for Focus/Background assignment (cf. Figure 4.10) to the quality or modifier type arguments of the verbal head models the following IS assignment rules.

IS Rule 33. *A non noun-modifier argument of the verbal head bearing the discourse information status of Rheme should also be assigned the information status of Focus.*

IS Rule 34. *A non noun-modifier argument of the verbal head bearing the discourse information status of Theme should be assigned the discourse information of Focus if the quality or property is to me made salient, otherwise Background.*

IS Rule 35. *A noun-modifier argument of the verbal head should be assigned the discourse information status of Focus if the noun referent is to be distinguished from other alternatives, otherwise Background.*

While the **IS Rule 33** provides for the rheme-focus marking for utterances such as (64a), **IS Rule 35** allows for the *marked* accent placement in (64b), and **IS Rule 34** provides for contrastive theme-focus marking, as shown in (64c).

(64) a. R: (Is the ball RED)$_{Rh}$
b. R: (Is that a RED ball?)$_{Rh}$
c. R: (The BALL)$_{Rh}$ (is RED?)$_{Th}$

The Focus/Background informational status of a referent is then specified by extending the proto-logical form with an informational feature. The leaf nodes in a decision tree, (cf. Figure 4.9 and 4.10), indicate the actions triggering the add-feature operation (see section 3.4) to specify the Focus information status using feature-value pair *Focus*=true and Background using the feature-value pair *Focus*=false.

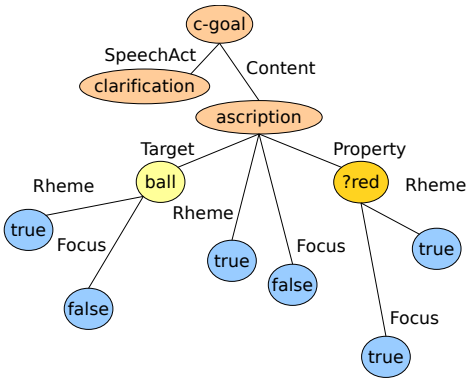


Figure 4.11: Extended proto-logical from with Focus/Background informativity status.

For example, the proto-logical form in Figure 4.8 is extended by the systems in Figure 4.9 and Figure 4.10 for specifying the Focus/Background information status for the subject and the predicate (the quality type nominal) in the proto-logical from. The extended relational structure for the proto-logical form is illustrated in Figure 4.11.

Encoding Agreement State as Information Structure

The decision tree in Figure 4.12 illustrates the *system* responsible for assigning the Agreement information status to entity and quality type nominal arguments of a verbal head of the sentence under planning. Based upon the general principles motivated in **Definition 3** the system models the following IS rules:

IS Rule 36. *A nominal bearing the information structure status of Focus, should be assigned the discourse information status of uncertain if the belief proposition corresponding to the referent is a partial knowledge or is void of information.*

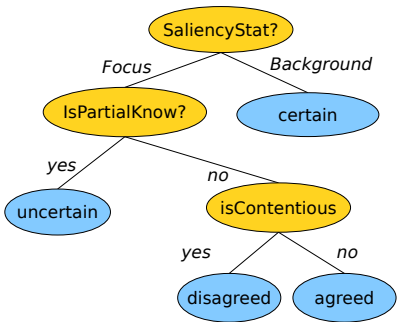


Figure 4.12: A system for assigning Agreement IS status.

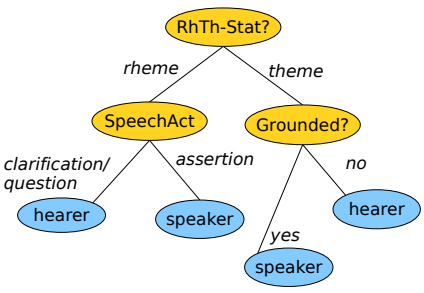


Figure 4.13: A system for assigning Commitment IS status.

IS Rule 37. *A event type nominal is assigned the default information status of certain.*

In our system partial or complete lack of knowledge in propositional content of privately acquired beliefs is represented in the beliefs propositions using the quantification operator ? (see section 3.3.1). The belief propositions in (65) illustrate the representation of uncertainty in a robot's belief propositions.

(65) a. $?t1.@_{\text{e1: entity}}(\mathsf{r}_{e1} \wedge \langle \mathit{Property} \rangle\,(\text{t1} : \mathit{type}))$

b. $?c1.@_{\text{e1: entity}}(\mathsf{r}_{e1} \wedge \langle \mathit{Property} \rangle\,(\text{c1} : \mathit{color} \wedge \mathsf{red}))$

The presence of quantified nominals in a proto-logical form enables the system in identifying the uncertainty or lack of knowledge, and map it in the Agreement dimension of the information structure by adding feature-value pair *Agreed*=plus to indicate certainty and feature-value pair *Agreed*=minus to indicate uncertainty.

Encoding Commitment State as Information Structure

The decision tree in Figure 4.13 illustrates the *system* responsible for assigning the Ownership information status to event, entity and quality type nominals in a proto-logical. Based upon the general principles motivated in **Definition 4** the system models the following IS rules.

IS Rule 38. *A nominal bearing the information structure status of Rheme, should be assigned the discourse information status of hearer owned when the communicative intention is to clarify or question.*

IS Rule 39. *A nominal bearing the information structure status of Rheme, should be assigned the discourse information status of speaker owned when the communicative intention is to make assertions or answering previously asked questions.*

IS Rule 40. *A nominal bearing the information structure status of Theme, should be assigned the discourse information status of speaker owned when the propositional content is grounded, otherwise, hearer owned.*

The informational status of a nominal with respect to the Commitment state is then specified by extending the proto-logical form with an informational feature. The leaf nodes in a decision tree (cf. Figure 4.13) indicate the actions triggering the add-feature operation (see section 3.4) to specify the information status using feature value *Owner*=hearer or *Owner*=speaker.

For example, the proto-logical form in Figure 4.11 is extended by the systems in Figure 4.12 and Figure 4.13 for specifying the Agreement and Ownership information status for the subject and the predicate (the quality type nominal) in the proto-logical from. The extended relational structure for the proto-logical form is illustrated in Figure 4.14.

Figure 4.15 provides a snapshot of the extensions to the planning grammar described in Figure 3.6 in section 3.4. While the relational structure in Figure

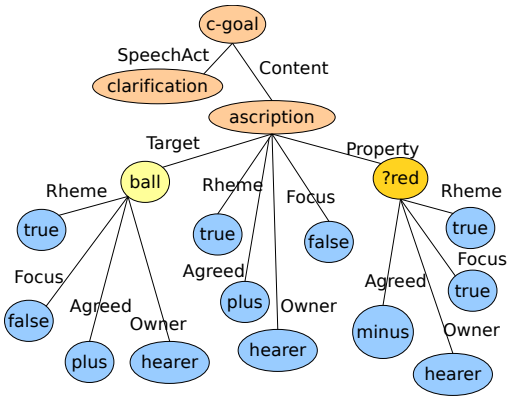


Figure 4.14: Extended proto-logical from with Agreement and Ownership informativity status.

4.14 indicates only the information structure feature-values for nominals sorts in the proto-logical form, various other systems in the planning grammar concurrently work for context sensitive utterance planning. The logical form in (66) illustrates the predicate-argument relations, semantic features encoding the contextual details and also the information structure semantics. This HLDS relational structure is realized into possible surface forms by the Surface Realizer component in the CogX which discuss in the next chapter.

(66) $@a1 : ascription(\text{be}$
$\wedge \langle Mood \rangle \text{interrogative}$
$\wedge \langle Tense \rangle \text{present}$
$\wedge \langle Rheme \rangle \text{true}$
$\wedge \langle Focus \rangle \text{false}$
$\wedge \langle Agreed \rangle \text{plus}$
$\wedge \langle Owner \rangle \text{hearer})$
$\wedge \langle Cop\text{-}Restr \rangle (e1 : entity \wedge \text{context}$
$\wedge \langle Delimitation \rangle \text{unique}$
$\wedge \langle Number \rangle \text{singular}$
$\wedge \langle Proximity \rangle \text{distal}$
$\wedge \langle Quantification \rangle \text{specific}$
$\wedge \langle Rheme \rangle \text{true}$
$\wedge \langle Focus \rangle \text{false}$
$\wedge \langle Agreed \rangle \text{plus}$
$\wedge \langle Owner \rangle \text{hearer})$
$\wedge \langle Cop\text{-}Scope \rangle (t1 : entity \wedge \text{box}$
$\wedge \langle Delimitation \rangle \text{existential}$
$\wedge \langle Number \rangle \text{singular}$
$\wedge \langle Quantification \rangle \text{specific}$
$\wedge \langle Rheme \rangle \text{true}$

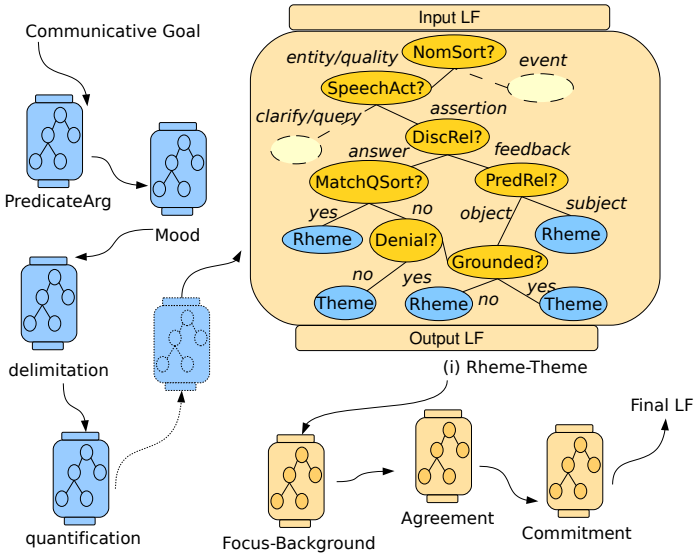


Figure 4.15: Systemic Grammar for Information Structure based utterance planning.

$$
\begin{aligned}
&\wedge \langle \mathit{Focus} \rangle \mathsf{false} \\
&\wedge \langle \mathit{Agreed} \rangle \mathsf{plus} \\
&\wedge \langle \mathit{Owner} \rangle \mathsf{hearer} \\
&\wedge \langle \mathit{Modifier} \rangle (c1 : \mathit{color} \wedge \mathsf{red} \\
&\qquad \wedge \langle \mathit{Rheme} \rangle \mathsf{true} \\
&\qquad \wedge \langle \mathit{Focus} \rangle \mathsf{true} \\
&\qquad \wedge \langle \mathit{Agreed} \rangle \mathsf{minus} \\
&\qquad \wedge \langle \mathit{Owner} \rangle \mathsf{hearer})) \\
\wedge \langle \mathit{Subject} \rangle &(e1 : \mathit{entity} \wedge \mathsf{context})
\end{aligned}
$$

4.4 Summary of the chapter

- In this chapter we motivated that the contextually appropriate intonational realization of robot utterances can be established by accounting for presenting (i) the informational state of the underlying beliefs with respect to the *common ground* established among the dialogue participants, (ii) the informational state of the underlying beliefs with respect to robot's *attentional state*? and (iv) a robot's attitude like *contentions* about the underlying beliefs, which may or may not have been yet *grounded*? and (iv) a robot's claim of *commitment* to the propositional content of the underlying beliefs?

  We have outlined a conceptual framework on how various informational aspects of an agent's beliefs, when represented as the four dimensions of the information structure, provide for the contextually appropriate presentation

of an utterance's meaning. Following this, we outlined the general principles for assigning information structure bearing in mind the robot's intentional and attentional state in dialogue context.

- We have presented a system implementation of the proposed approach. We have illustrated our approach to gather these four informational aspects of an agent's beliefs, and representing them as part of the linguistic meaning of a communicative goal. As a result of this exercise we obtain a common relational structure representing both the linguistic meaning and its information structure partitions. Such a representation allows as to derive the surface from and its intonation contour using a grammar framework that models the information structure semantics. We will discuss the modeling of such a grammar in the following chapter.

- While the contextual framework presented here addresses a diverse range of presentational styles for a robots belief state, the implemented system covers a rather small set of these phenomenon. For example, we discussed that a speaker presupposes a hearer to accommodate the claims made by the speaker about the hearer's knowledge. We still have to see when and why a robot would intend to make such a presupposition. The current implementation of the system relies on the actual state of common ground for such a presentation, thereby leaving out such an usage.

- A conversation may also involve disagreement over the propositions contributed by the participants to the context. Although we have motivated that an agent's agreement and disagreement to contributed information can be modeled in the same lines as modeling uncertainty, the current implementation of the system mostly grounds the information conveyed to it by the tutor. This is primarily due to the fact that the system's learning abilities are still primitive. However, as the system evolves in its learning abilities, presentation of disagreement about grounding attributed beliefs would prove to be an important feature in a robot's dialogue abilities.

5 Modeling Intonation

In this chapter, we describe our approach to the intonational realization of information structure. We show how a semantic representation containing the *predicate-argument relation* and *information structure* details is realized into its surface form. We start with a brief discussion on Steedman's model of *combinatory prosody*. Next, using this model, we present our account of modeling information structure and intonation in the CCG framework. After this we elaborate on the finer aspects of the grammar implementation on the OpenCCG platform. We describe how *pitch accents*, *boundary tones*, the rules governing prosodic derivations and compositional intonational semantics are specified in the grammar. We illustrate a variety of prosodic derivations that our grammar supports. We end with a discussion on issues that arise in implementing some of the theoretical aspects of combinatory prosody.

5.1 Realizing Intonation

Intonation in spoken English provides the contextual framework for analyzing the meaning of an utterance; for relating the meaning of utterances to one another; and for enabling a listener to access the speaker's intentions and attentional state. Each of these functions of intonation is conveyed by a small set of intonational parameters which indicate *phrasing* and *accentual patterns*. The ability of a robot to produce utterances with contextually appropriate intonation should therefore allow it to rightly express its intentions, thereby enhancing the interpretability of its utterances, and their coherence in a discourse context.

In order to enable a robot to produce utterances with intonation we need to simulate an understanding of the significance of the melodies of spoken language. The theory of compositional intonation proposed by Pierrehumbert and Hirschberg [1990], suggests that the phrasal and accentual patterns are compositional in their meaning – composed from the *pitch accents*, *phrase accents*, and *boundary tones*. Their theory provides us with an understanding of the meaning of intonational tunes and their contribution to discourse interpretation. However, what is required for realizing intonation in utterances is a framework in which the discourse semantics

of an utterance is mapped to the intonational semantics of tunes.

Steedman [2000b,a] (see section 2.2.1 and 2.3) offers a framework which (*i*) associates intonation with discourse meaning in terms of information structure (IS), (*ii*) provides a compositional semantics of English intonation in information-structural terms, (*iii*) tightly couples intonation with grammatical structure, and (*iv*) assumes a general IS-sensitive notion of discourse context update. Modeling information structure is therefore key to the realizing of intonation in robot utterances.

In Chapter 4 we have elaborated our approach to model Steedman's four dimensions of information structure. We discussed our approach for capturing the informational aspects of a discourse – *Theme/Rheme*, *Focus/Background*, and agent attitudes like *Agreement* and *Commitment*. These aspects are then encoded as the four dimensions of informativity in the linguistic meaning (predicate-argument relations) representation of an utterance. As an end result, we obtain one common relational structure representing an utterance's linguistic meaning and its information structure. Both these pieces of information are vital for the realizations of the surface form and its intonational contour.

The surface form and the intonational realization of the linguistic meaning of an utterance requires a grammar that models the compositional aspects of syntactic, semantic and intonational meaning of spoken English. In this chapter we describe our approach to modeling intonation in a grammar following Steedman's [2000a] model of *combinatory prosody*. We start with a brief discussion of this model. Following this we elaborate on our approach and implementation of this model in the OpenCCG framework [Baldridge, 2002; Baldridge and Kruijff, 2002] of CCG [Steedman, 2000b].

combinatory prosody

5.1.1 Intonation and Information Structure

In Steedman's [2000a] model of combinatory prosody, *pitch accents* and *boundary tones* have an effect on both the syntactic category of the expression they mark, as well as the meaning of that expression. Following the discussions in [Pierrehumbert and Beckman, 1986] over the categorization of pitch accents, Steedman distinguishes pitch accents as markers of either the *theme* (θ) or of the *rheme* (ρ) information structure units. Accordingly, pitch accents tones L+H* and L*+H are referred to as ***θ-markers***, whereas H*, L*, H*+L and H+L* tones are referred to as ***ρ-markers***.

θ-marking
ρ-marking

Since pitch accents mark individual words and not (necessarily) larger phrases, Steedman uses the *θ/ρ-marking* to spread informativity over the *domain* and the *range* of the function categories. Therefore, the identical markings on different parts of a functional category act not only as features, but also as occurrences of a singular variable. The value of the marking on the domain can thus get passed down ("projected") to markings on categories in the range. In the model, individual words bearing no pitch accent are referred to as ***η-marked*** categories. The η-marked categories can further unify with either η, θ or ρ marked categories.

η-marking

A θ/ρ-marked category may then combine with either categories bearing similar pitch accent markings and form larger marked phrases, or with the *phrasal*

tones L or H, and result in an *intermediate phrase* (in terms of [Pierrehumbert and Hirschberg, 1990]). An intermediate phrase is, however, "incomplete" until it combines with the boundary tones L% or H% and results into an *intonational phrase*. Boundary tones, have thus the effect of mapping phrasal tones into intonational phrase boundaries. An intonational phrase can then combine with other complete intonational or intermediate phrases. To make these boundaries explicit, and enforce "complete" prosodic phrases to combine with other complete prosodic phrases, Steedman introduces two further types of markings: ι and ϕ on categories. These markings are, however, purely syntactic in nature, and as such does not affect the θ/ρ informativity of the categories.

An ***ι-marked*** category is a pitch accent bearing category that has combined with a phrasal tone (either L or H) to its right. That is, a ι-marked category basically represents an intermediate phrase. A ***ϕ-marked*** category, on the other hand is an intermediate phrase that has combined with a boundary tone (either L% or H%) to its right, resulting into a "complete" intonational phrase. Thus a ϕ-marked category is basically an intonational phrase. A ϕ-marked phrases can unify with only other ϕ or ι-marked phrases, not with η, θ or ρ marked categories. The ϕ or ι-markings are introduced only to provide derivational control and are not reflected in the underlying meaning (which is reflected by η, θ or ρ-markings).

ι-marking

ϕ-marking

Intonational phrasal units are thus compositionally built in the same manner as phrases in a sentence are built from combination of word categories. Furthermore, as with the sentence semantics which is compositional, the meaning contained in intonational phrases is also compositional, and is driven by the syntactic derivations under the constraints of prosodic combinatory. Modeling combinatory prosody in a grammar framework therefore requires one to specify (i) the constraints of the prosodic derivations alongwith the constraints governing the syntactic derivations of the underlying language, and (ii) the compositional semantics of intonational tunes as the information structure of the utterance meaning.

Contemporary unification based grammar frameworks use *sign* as a means for representing the linguistic constructs, such as word categories, in the grammar. These signs are multi-dimensional, with each dimension (or level) representing an aspect of linguistic information such as *syntax*, *semantics*, *phonology*, etc. Using multi-level signs, it is possible to represent information of various domains, and allow these domains to interact with each other and help constrain the unification process. What we need for our implementation is a sign that is also able to represent the *prosodic* and *information structure* dimensions along with the syntactic, semantic and phonology dimensions, and also provide for modeling the constraints of combinatory prosody.

sign

In this direction, we use the generalized multi-level CCG signs described in Kruijff and Baldridge [2004]. The following section elaborates on our approach to employing these *multi-level signs* for modeling combinatory prosody in a CCG framework. We will illustrate how the θ, ρ, *η-marking* of Steedman's model can be specified in multi-level signs, and how derivational constraints corresponding to ι and *ϕ-marking* can be achieved. After this, in section 5.3 we describe the

multi-level signs

implementation details of this approach on the OpenCCG platform for grammar development.

5.2 Multi-level Signs in CCG

A sign is an n-tuple of terms that represent information in n distinct dimensions. Each dimension represents a level of linguistic information, such as prosody, meaning, or syntactic category. As a representation, we assume that for each dimension we have a language that defines well-formed representations, and a set of operations which can create new representations from a set of given representations.

Kruijff and Baldridge [2004] define a multi-level sign with the following five dimensions:

- **Phonology**: representing word or word sequences, and composition of sequences takes place through concatenation.
- **Prosody**: representing tunes or sequences of tunes from the inventory of intonational theory of [Pierrehumbert and Hirschberg, 1990]. Composition of tunes sequence takes place through concatenation.
- **Syntactic Category**: well-formed CCG categories, and combinatory rules (see section 3.5).
- **Information structure**: a hybrid logic formula of the form $@_d[in]\mathsf{r}$, where r corresponds to a discourse referent that has informativity *in*, theme (θ), or rheme (ρ) relative to the current point in the discourse d [Kruijff, 2003].
- **Predicate-argument structure**: hybrid logic formulas of the form discussed in section 3.2, representing word or phrasal semantics.

Example (67) illustrates a sign with these five dimensions. The word-form box bears the H* pitch accent, and acts as a noun type syntactic category n. The pitch accent H* indicates that the discourse referent m (from n_m) introduces *new* information at the current point in the discourse d. This suggests that the meaning $@_m\mathsf{box}$ should end up as part of the rheme (ρ) information unit of the utterance. This detail is specified in the information structure dimension as $@_d[\rho]\mathsf{m}$.

(67)
$$\begin{array}{c} \mathsf{BOX} \\ \mathsf{H^*} \\ \hline \mathsf{n}_m \\ @_d\,[\rho]\,\mathsf{m} \\ @_m\mathsf{box} \end{array}$$

If a sign does not specify any information at a particular dimension, this is indicated by $\top$ (or an empty line). Example (67) represents a pitch accent marked category, more specifically a ***ρ-marked*** category. It thereby indicates that the word belongs to the rheme information unit of an utterance. On the other hand, the sign in (68) represents a ***θ-marked*** category, indicating that it should belong to the theme information unit of an utterance.

(68)
$$\frac{\begin{array}{c}\textsf{GREEN}\\ \textsf{L+H*}\end{array}}{\begin{array}{c}\textsf{adj}_r\\ @_d\,[\theta]\,\textsf{r}\\ @_r\textsf{green}\end{array}}$$

The word-form **green** with the L+H* pitch accent in (68) acts as a modifier (or adjunct) category. The L+H* accent indicates that the discourse referent r (from $\textsf{adj}_r$) introduces a piece of information that has either been previously mentioned (or given) or is presupposed by the speaker at the current point in the discourse d. It thereby indicates that the meaning $@_r\textsf{green}$ should end up as part of the theme (θ) information unit of the utterance. This is rightly indicated by the information structure dimension, $@_d[\theta]\textsf{r}$.

In order to elaborate the unifications of multi-level signs under prosodic constraints, we use as an example the human- robot interaction in Figure 5.1. The dialogue fragment in (69) illustrates the respective turns.

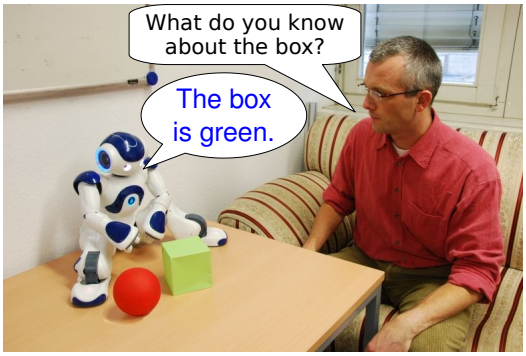


Figure 5.1: Contextual setup for the dialogue in (69)

(69) a. H: Robot, what do you know about the box?

b. R: (The BOX)$_{Th}$ (is GREEN)$_{Rh}$
L+H* LH% H* LL%

Example (70) illustrates all the signs corresponding to the content words of the robot's utterance in (69b).

(70)
$$\frac{\begin{array}{c}\textsf{The}\\ \top\end{array}}{\begin{array}{c}\textsf{s}_i/(\textsf{s}_i\backslash!\textsf{np}_i)/\textsf{n}_t\\ @_d\,[\eta]\,\textsf{t}\\ @_t\textbf{the}\wedge @_t\langle\textsc{n}\rangle u\end{array}}\quad\frac{\begin{array}{c}\textsf{BOX}\\ \textsf{L+H*}\end{array}}{\begin{array}{c}\textsf{n}_m\\ @_d\,[\theta]\,\textsf{m}\\ @_m\textbf{box}\end{array}}\quad\frac{\begin{array}{c}\textsf{is}\\ \top\end{array}}{\begin{array}{c}\textsf{s}_i\backslash!\textsf{np}_x/\textsf{adj}_y\\ @_d\,[\eta]\,\textsf{v}\\ @_v\textbf{is}\wedge @_v\langle\textsc{act}\rangle x\wedge @_v\langle\textsc{pat}\rangle y\end{array}}\quad\frac{\begin{array}{c}\textsf{GREEN}\\ \textsf{H*}\end{array}}{\begin{array}{c}\textsf{adj}_r\\ @_d\,[\rho]\,\textsf{r}\\ @_r\textbf{green}\end{array}}$$

Observe that the information structure dimension of the signs for words 'box' and 'green' bear Steedman's θ and *ρ-marking* respectively. On the other hand, as the words 'The' and 'is' in (69b) do not bear any pitch accent, the prosodic dimension

of the corresponding signs in (70) contain a $\top$ and their information structure dimension bears Steedman's *η-marking*.

unification

At this point, an *η-marked* sign can unify with any other η or θ/ρ-marked signs and result into larger phrases. During the *unification*, the combinatory operations at the phonology and the prosodic dimension are simple string concatenations. In the semantic dimension, the HLDS relations structure is extended as detailed in section 3.2. In a similar way, the HLDS relational structure in the information structure dimension is extended. However, the informativity status of the unmarked categories is now updated with the informativity status of the marked categories. In this manner, the informativity of marked words is spread over the unmarked constituents of an information structure unit.

The unification of signs, however, is driven by the combinatory rules working at the syntactic dimension. The syntactic combinations of the signs in (70) takes place as illustrated in example (71).

(71)

$$\dfrac{\dfrac{\begin{array}{c}\mathsf{The}\\ \top\end{array}}{\begin{array}{c}\mathsf{s}_i/(\mathsf{s}_i\backslash!\mathsf{np}_i)/\mathsf{n}_t\\ @_d\,[\eta]\,\mathsf{t}\\ @_t\mathbf{the}\wedge @_t\langle\mathsf{n}\rangle u\end{array}} \quad \dfrac{\begin{array}{c}\mathsf{BOX}\\ \mathsf{L{+}H^*}\end{array}}{\begin{array}{c}\mathsf{n}_m\\ @_d\,[\theta]\,\mathsf{m}\\ @_m\mathbf{box}\end{array}}}{\begin{array}{c}\mathsf{s}_i/(\mathsf{s}_i\backslash!\mathsf{np}_i)\\ @_d\,[\theta]\,\mathsf{t}\wedge @_d\,[\theta]\,\mathsf{m}\\ @_t\mathbf{the}\wedge @_t\langle\textsc{n}\rangle m\wedge @_m\mathbf{box}\end{array}}>$$

$$\dfrac{\dfrac{\begin{array}{c}\mathsf{is}\\ \top\end{array}}{\begin{array}{c}\mathsf{s}_i\backslash!\mathsf{np}_x/\mathsf{adj}_y\\ @_d\,[\eta]\,\mathsf{v}\\ @_v\mathbf{is}\wedge @_v\langle\textsc{act}\rangle x\wedge @_v\langle\textsc{pat}\rangle y\end{array}} \quad \dfrac{\begin{array}{c}\mathsf{GREEN}\\ \mathsf{H^*}\end{array}}{\begin{array}{c}\mathsf{adj}_r\\ @_d\,[\rho]\,\mathsf{r}\\ @_r\mathbf{green}\end{array}}}{\begin{array}{c}\mathsf{s}_p\backslash!\mathsf{np}_x\\ @_d\,[\rho]\,\mathsf{v}\wedge @_d\,[\rho]\,\mathsf{r}\\ @_v\mathbf{is}\wedge @_v\langle\textsc{act}\rangle x\wedge @_v\langle\textsc{pat}\rangle r\wedge @_r\mathbf{green}\end{array}}>$$

The signs for the determiner 'The' and the noun 'box' unify as a consequence of forward operation (>) and result into the noun phrase "The BOX". Similarly, the sign for verbal head 'is' seeking a adjective to its right, unifies with the modifier 'green' and results into the verb phrase "is GREEN". The phonology dimension of the resulting categories indicate these phrases.

Note that the modal operator ! on the slash operator in category $\mathsf{s}_i\backslash!\mathsf{np}_x/\mathsf{adj}_y$ prevents the verbal head to first combine with a np category to its left. The presence of the ! operator makes the otherwise functional category $\mathsf{s}_i\backslash!\mathsf{np}_x$ behave as an atomic category.

In the syntactic dimension we are left with the categories from the range of the functional categories for words 'The' and 'is'. In the semantic dimension, semantic unifications constrained by the sorted indices result in extended HLDS relation structures, which respectively indicate the predicate-argument structure of the resulting phrases.

Of prime significance to us at this point is the information structure dimension of the resultant categories. One can observe that in (71) the informativity θ and ρ of pitch accent marked words 'box' and 'green' respectively has been spread over the corresponding phrases "The BOX" and "is GREEN". We can infer from the informativity status of these phrasal categories that they belong to the theme and rheme units of information structure of the sentence under derivation.

From the perspective of Steedman's model, the unification process which has resulted in the formation of these information structure units should also provide for their respective intonational contours. The prosodic dimension of the resulting categories in (71) illustrate the intonational tunes corresponding to these IS units. The pitch accents L+H* and H* in the intonational contour align with the words in SMALLCAPS in the phrases "The BOX" and "is GREEN" respectively.

The intonational tunes marking the information structure units at this stage are "incomplete". Any further unification of these signs at the syntactic level needs to be constrained with the fact that categories with only "complete" theme and rheme intonational phrases (at the prosodic dimension) can unify. What we mean with this is that marked phrases should first unify with phrasal tones L or H and result into an intermediate phrase. In Steedman's model, θ/ρ-marked categories with a phrasal tone on their left are referred to as *ι-marked* categories. An intermediate phrases may then further combine with either another *ι-marked* category or with a boundary tone (L% or H%). This results in a complete intonational phrases, which is referred to as a *ϕ-marked* category in Steedman's model.

So how do we model phrasal and boundary tones using multi-level signs? Since the role of the phrasal and the boundary tones in the intonational phrases is similar to that of the punctuation marks in a sentence, we could model phrasal and boundary tones as just another string element in the lexicon. Since they are used for *derivational control* we need to assign them syntactic categories.

Keeping the unification constraints of phrasal and boundary tone in view, we define phrasal tones as categories which take a ρ or *θ-marked* phrasal categories as argument and result into the same phrasal category, albeit with an *ι-marking*. Example (72) illustrates a multi-level sign corresponding to phrasal tone L. Observe that for phrasal tones, all other dimensions are $\top$ except for the syntactic and the prosodic dimesions. The signs for phrasal tone L and H differ only in their prosodic dimension.

(72) $$\frac{\mathsf{L}}{\mathsf{s\$}_{\iota}\backslash_{\star}\mathsf{s\$}_{m}}$$

The \$ notation in domain and range of the functional category in (72) indicates that these categories could be anything (atomic or functional categories) with resulting category as a s. The subscript m in the syntactic dimension indicates that the argument to this category is a pitch accent marked phrase, whereas the subscript ι indicates that the resultant phrasal category is *ι-marked*, showing that it is an intermediate phrase.

The syntactic derivations in example (73) illustrate the composition of an intermediate phrase where the phrasal category for "The BOX" unifies with phrasal tone L. The resultant category indicates that phrasal tones affect the syntactic category with an ι-marking, as in $\mathsf{s}_{\iota}/(\mathsf{s}_{\iota}\backslash!\mathsf{np}_{i})$, indicating that the phrasal category is an intermediate phrase. This allows us to achieve the further derivational control in the prosodic dimension where we now have the intermediate phrase "L+H* L".

(73)

$$
\dfrac{
\dfrac{
\dfrac{\text{The} \quad \top}{\begin{array}{c} \mathsf{s}_i/(\mathsf{s}_i\backslash !\mathsf{np}_i)/\mathsf{n}_t \\ @_d\,[\eta]\,\mathsf{t} \\ @_t\textbf{the} \wedge @_t\langle \mathrm{N}\rangle u \end{array}}
\quad
\dfrac{\text{BOX} \quad \text{L+H*}}{\begin{array}{c} \mathsf{n}_m \\ @_d\,[\theta]\,\mathsf{m} \\ @_m\textbf{box} \end{array}}
}{\begin{array}{c} \mathsf{s}_i/(\mathsf{s}_i\backslash !\mathsf{np}_i) \\ @_d\,[\theta]\,\mathsf{t} \wedge @_d\,[\theta]\,\mathsf{m} \\ @_t\textbf{the} \wedge @_t\langle \mathrm{N}\rangle m \wedge @_m\textbf{box} \end{array}}>
\quad
\dfrac{\text{L}}{\mathsf{s}\$_\iota\backslash_\star \mathsf{s}\$_m}
}{\begin{array}{c} \mathsf{s}_\iota/(\mathsf{s}_\iota\backslash !\mathsf{np}_i) \\ @_d\,[\theta]\,\mathsf{t} \wedge @_d\,[\theta]\,\mathsf{m} \\ @_t\textbf{the} \wedge @_t\langle \mathrm{N}\rangle m \wedge @_m\textbf{box} \end{array}}<
$$

At this point, according to Steedman's model an *ι-marked* phrase can combine with either another ι-marked phrase or with the boundary tones L% or H% and result into a "complete" intonational phrase i.e a *ϕ-marked*. The purpose of ϕ-marking is also to provide further derivational control and is not reflected in the underlying meaning. Similar to phrasal tones, we define boundary tones as strings in the lexicon with the multi-level sign representation as in (74), where all other dimensions being $\top$ except for the prosodic and the syntactic dimesions.

(74)

$$
\dfrac{\text{H\%}}{\mathsf{s}\$_\phi\backslash_\star \mathsf{s}\$_\iota}
$$

The subscripts on the syntactic categories indicate that a boundary tone takes a ι-marked phrasal category as argument and results in the same phrasal category but with ϕ-marking. The signs for boundary tones H% and L% differ only in their prosodic dimension.

Driven by the operations at the syntactic dimension, in the next step the sign for boundary tone H% in (74) unifies with the ι-marked category in (73) and result in an intonational phrase i.e. a *ϕ-marked* phrasal category. Example (75) illustrates this derivation.

(75)

$$
\dfrac{
\dfrac{
\dfrac{
\dfrac{\text{The} \quad \top}{\begin{array}{c} \mathsf{s}_i/(\mathsf{s}_i\backslash !\mathsf{np}_i)/\mathsf{n}_t \\ @_d\,[\eta]\,\mathsf{t} \\ @_t\textbf{the} \wedge @_t\langle \mathrm{N}\rangle u \end{array}}
\quad
\dfrac{\text{BOX} \quad \text{L+H*}}{\begin{array}{c} \mathsf{n}_m \\ @_d\,[\theta]\,\mathsf{m} \\ @_m\textbf{box} \end{array}}
}{\begin{array}{c} \mathsf{s}_i/(\mathsf{s}_i\backslash !\mathsf{np}_i) \\ @_d\,[\theta]\,\mathsf{t} \wedge @_d\,[\theta]\,\mathsf{m} \\ @_t\textbf{the} \wedge @_t\langle \mathrm{N}\rangle m \wedge @_m\textbf{box} \end{array}}>
\quad
\dfrac{\text{L}}{\mathsf{s}\$_\iota\backslash_\star \mathsf{s}\$_m}
}{\begin{array}{c} \mathsf{s}_\iota/(\mathsf{s}_\iota\backslash !\mathsf{np}_i) \\ @_d\,[\theta]\,\mathsf{t} \wedge @_d\,[\theta]\,\mathsf{m} \\ @_t\textbf{the} \wedge @_t\langle \mathrm{N}\rangle m \wedge @_m\textbf{box} \end{array}}<
\quad
\dfrac{\text{H\%}}{\mathsf{s}\$_\phi\backslash_\star \mathsf{s}\$_\iota}
}{\begin{array}{c} \mathsf{s}_\phi/(\mathsf{s}_\phi\backslash !\mathsf{np}_i) \\ @_d\,[\theta]\,\mathsf{t} \wedge @_d\,[\theta]\,\mathsf{m} \\ @_t\textbf{the} \wedge @_t\langle \mathrm{N}\rangle m \wedge @_m\textbf{box} \end{array}}<
$$

Lets take a closer look at the resultant multi-level sign in (75). At the phonology dimension we still have the same word sequence "The BOX". The dimension for information structure with a *θ-marking* indicates that the phrase bears the theme informativity. In the prosodic dimension we now have a complete intonational phrase "L+H* LH%", which corresponds to the intonation tune for a theme information unit in Steedman's model.

In a similar manner, the phrasal category for "is GREEN" first combines with phrasal tone L and then with boundary tone L% and result into another complete intonation phrase. Example (76) illustrates this derivation.

(76)

$$\begin{array}{cccc}
\dfrac{\begin{array}{c}\mathsf{is}\\ \top\end{array}}{\begin{array}{c}\mathsf{s}_i\backslash!\mathsf{np}_x/\mathsf{adj}_y\\ @_d\,[\eta]\,\mathsf{v}\\ @_v\textbf{is}\wedge @_v\langle\text{ACT}\rangle x\wedge @_v\langle\text{PAT}\rangle y\end{array}} &
\dfrac{\begin{array}{c}\mathsf{GREEN}\\ \mathsf{H*}\end{array}}{\begin{array}{c}\mathsf{adj}_r\\ @_d\,[\rho]\,\mathsf{r}\\ @_r\textbf{green}\end{array}} &
\dfrac{\mathsf{L}}{\mathsf{s}\$_\iota\backslash_\star\mathsf{s}\$_i} &
\dfrac{\mathsf{L\%}}{\mathsf{s}\$_\phi\backslash_\star\mathsf{s}\$_\iota}
\end{array}$$

$$\xrightarrow{>}\quad \begin{array}{c}\mathsf{s}_i\backslash!\mathsf{np}_x\\ @_d\,[\rho]\,\mathsf{v}\wedge @_d\,[\rho]\,\mathsf{r}\\ @_v\textbf{is}\wedge @_v\langle\text{ACT}\rangle x\wedge @_v\langle\text{PAT}\rangle r\wedge @_r\textbf{green}\end{array}$$

$$\xrightarrow{<}\quad \begin{array}{c}\mathsf{s}_\iota\backslash!\mathsf{np}_x\\ @_d\,[\rho]\,\mathsf{v}\wedge @_d\,[\rho]\,\mathsf{r}\\ @_v\textbf{is}\wedge @_v\langle\text{ACT}\rangle x\wedge @_v\langle\text{PAT}\rangle r\wedge @_r\textbf{green}\end{array}$$

$$\xrightarrow{<}\quad \begin{array}{c}\mathsf{s}_\phi\backslash!\mathsf{np}_x\\ @_d\,[\rho]\,\mathsf{v}\wedge @_d\,[\rho]\,\mathsf{r}\\ @_v\textbf{is}\wedge @_v\langle\text{ACT}\rangle x\wedge @_v\langle\text{PAT}\rangle r\wedge @_r\textbf{green}\end{array}$$

The resulting intonational phrase "H* LL%" in the prosodic dimension corresponds to the rheme intonational tune in Steedman's model. This further indicates that the marked phrase bears the rheme informativity. The *ρ-marking* in the information structure dimension of the resultant category confirms this.

Observing the syntactic categories of the resultant phrasal categories in (75) and (76) we note that the former seeks the latter; they are both *ϕ-marked* (complete intonational phrases), and therefore can unify at the syntactic and the prosodic dimensions. Derivation in (77) illustrates the combinations on the respective dimensions.

(77)

$$\begin{array}{cc}
\dfrac{\begin{array}{c}\mathsf{The\ BOX}\\ \mathsf{L{+}H*\quad LH\%}\end{array}}{\begin{array}{c}\mathsf{s}_\phi/(\mathsf{s}_\phi\backslash!\mathsf{np}_i)\\ @_d\,[\theta]\,\mathsf{t}\wedge @_d\,[\theta]\,\mathsf{m}\\ @_t\textbf{the}\wedge @_t\langle\text{N}\rangle m\wedge @_m\textbf{box}\end{array}} &
\dfrac{\begin{array}{c}\mathsf{is\ GREEN}\\ \mathsf{H*\quad LL\%}\end{array}}{\begin{array}{c}\mathsf{s}_\phi\backslash!\mathsf{np}_x\\ @_d\,[\rho]\,\mathsf{v}\wedge @_d\,[\rho]\,\mathsf{r}\\ @_v\textbf{is}\wedge @_v\langle\text{ACT}\rangle x\wedge @_v\langle\text{PAT}\rangle r\wedge @_r\textbf{green}\end{array}}
\end{array}$$

$$\xrightarrow{<}\quad \begin{array}{c}\mathsf{s}_\phi\\ @_d\,[\theta]\,\mathsf{t}\wedge @_d\,[\theta]\,\mathsf{m}\wedge @_d\,[\rho]\,\mathsf{v}\wedge @_d\,[\rho]\,\mathsf{r}\\ @_v\textbf{is}\wedge @_v\langle\text{ACT}\rangle t\wedge @_v\langle\text{PAT}\rangle r\wedge @_r\textbf{green}\wedge @_t\textbf{the}\wedge @_t\langle\text{N}\rangle m\wedge @_m\textbf{box}\end{array}$$

The derivation in (77) results into a complete sentence, as can be observed from the category s_ϕ in the syntactic dimension. The phonology dimension contains the surface form – the sentence, "The BOX is GREEN". The prosody dimension indicates

the sentence's intonational contour, "L+H* LH% H* LL%", where the pitch accents align with the words they mark in the phonological string. The boundary tones LH% and LL% mark the theme and rheme intonational phrases boundaries. This suggests that the sentence is composed of two information units. The information structure dimension indicates the theme informativity status of discourse referents m and t, and the rheme informativity status of referents r and v in the discourse d. Finally the semantic dimension indicates the underlying predicate-argument relation that resulted from the derivations leading to the realization of the surface form.

In Chapter 4, we have elaborated on our approach to modeling information structure in an utterance's linguistic meaning. Remember that we encode the information structure and the predicate-argument relation using the HLDS representations. Using the multi-level signs of Kruijff and Baldridge [2004] we have just shown how the derivations of Steedman's [2000a] model of combinatory prosody can be implemented in the CCG framework. We illustrated that, just as the sentence meaning is compositionally built at the semantic dimension during the syntactic combinations of individual word categories, the intonational meaning is built compositionally and concurrently in the information structure dimension during the prosodic derivations of the pitch accents and the boundary tones. Since such a grammar can also be used for *realization*, the semantics and information structure representations of Chapter 4 can be easily realized into surface forms containing intonational contours.

realization

In the following section we discuss the finer aspects of grammar implementation of the multi-level signs, the various marking and prosodic constraints, in the OpenCCG platform for CCG. Crucial to the development of such a grammar is the notion of *sign* and the various sorts of *markings* proposed in Steedman's model.

5.3 Implementing a Prosodic Grammar

Any grammar framework for the purpose of natural language generation and parsing, requires a means of representing the *linguistic constructs* and the *grammar rules* that govern combinations of these constructs into valid structures of the underlying language. The means for representation of linguistic constructs in OpenCCG is a three-dimensional *sign*, (see section 3.5), represented as follows:

(78) phonology $\vdash$ syntax : semantics

The left hand side of the $\vdash$ operator in (78) is the phonology dimension, which contains the word form(s). The right hand side of the operator indicates the syntactic and semantic dimensions. The : operator separates these two dimensions. What this signifies is that the semantics of the unifying categories is compositional, and is governed by the syntactic derivations. Example (79) is a sign representing the *lexical entry* for a noun type category 'box'.

lexical entry

(79) box $\vdash$ n : $@_{m:sort}(\mathbf{box})$

In the previous section, while developing an approach for modeling combinatory prosody in CCG, we used a multi-level sign with the following five dimensions: phonology, prosody, syntactic category, information structure and semantics. The sign in (80) summarizes this formulation for the noun type entity 'box', bearing the pitch accent H* i.e. a *ρ-marked* category.

(80)
$$\frac{\begin{array}{c}\mathsf{BOX}\\ \mathsf{H^*}\end{array}}{\begin{array}{c}\mathsf{n}_m\\ @_d\,[\rho]\,\mathsf{m}\\ @_m\mathbf{box}\end{array}}$$

Given this five-dimensional sign and three-dimensional sign of OpenCCG, the first question that we need to address towards grammar implementation is: *how do we represent the prosody and information structure dimensions of the five-dimensional sign in the* OpenCCG *framework*? The question that we need to address next is: *how do we model the constraints that govern the unifications at the prosody dimension*? Let us start with the first question.

Observe that the combinatory operations governing the unifications at the phonology and the prosody dimensions are simple string concatenation. That is, tunes combine to form a sequence of tunes in the same manner as words combine to form a sequence of words. We can therefore collapse these two dimensions of the five-dimensional sign together. The four-dimensional sign in (81) illustrates the collapsing of the phonology and prosody dimensions of the corresponding five-dimensional sign in (80).

(81)
$$\frac{\mathsf{BOX\_H^*}}{\begin{array}{c}\mathsf{n}_m\\ @_d\,[\rho]\,\mathsf{m}\\ @_m\mathbf{box}\end{array}}$$

This implies that in the grammar lexicon, pitch accent marked words are represented as a combination of their phonological form and the pitch accent tune.

Observe next that the logic used for representation of information at the semantic and information structure dimensions is HLDS. Therefore we can again collapse these two dimensions of the four-dimensional sign. The three-dimensional sign in (82) illustrates the collapsing of the semantic and the information structure dimensions of the four-dimensional sign in (81).

(82)
$$\frac{\mathsf{BOX\_H^*}}{\begin{array}{c}\mathsf{n}_m\\ @_m\mathbf{box} \wedge @_d\,[\rho]\,\mathsf{m}\end{array}}$$

It is relatively obvious from this collapsed representation that the semantic and the information structure features indicate the contextual and informational state aspect of the discourse entity m at a given point in a discourse d.

With these two modifications we are able to reduce the five-dimensional sign into a three-dimensional sign, which is compatible with the lexicon sign in OpenCCG.

Marking	Pitch Accents	Example Lexicon
η-marking	nulltone	box, green
ρ-marking	H*, L*, H*+L, H+L*	box_H*, green_H*, box_L*, green_L*
θ-marking	L+H*, L*+H	box_L+H*, green_L+H*

Table 5.1: Lexicon representation of prosodic markings.

The approach to reduction of dimensions is taken to exploit the CCG combinatorics already available with the OpenCCG framework. Alternatively, one could think of modifying the OpenCCG implementation for extending the three-dimensional signs to a five-dimensional sign, but this is altogetherly out of the scope of our current work. Besides that, the approach taken in this work illustrate that the prosodic realizations can be derived using the combinatorics that operates on surface realizations.

We can therefore modify the representation of lexicon entries in OpenCCG as the following:

(83) $\mathsf{phonology\_prosody} \vdash \mathsf{syntax} : \mathsf{semantics} \wedge \mathsf{information\ structure}$

So the phonology dimension now contains the concatenation of the word form and prosodic tunes that marks it. We reserve the symbol _ to mark the separation of the phonological and the prosodic forms. Following this, the lexical entry corresponding to the pitch accent marked sign in (80) takes the following form in (84). The lexical entry for an unmarked category such as (85), is given as in (86).

(84) $\mathsf{BOX\_H*} \vdash \mathsf{n} : @_{m:sort}(\mathbf{box}) \wedge @_d\,[\rho]\,\mathsf{m}$

(85)
$$\frac{\mathsf{box}}{\mathsf{n}_m}$$
$$@_m\mathbf{box} \wedge @_d\,[\eta]\,\mathsf{m}$$

(86) $\mathsf{box} \vdash \mathsf{n} : @_{m:sort}(\mathbf{box}) \wedge @_d\,[\eta]\,\mathsf{m}$

However, due to the perspicuousness of the collapsed multi-level sign notation, such as those in (82) or (85), we continue to use them in the rest of this chapter for illustration the of the prosodic derivations, instead of using the OpenCCG lexicon entries like (84) or (86).

Having addressed the first question on the lexical representation of a five-dimensional sign, let us take a closer look at what exactly the information structure terms $@_d\,[\rho]\,\mathsf{m}$ and $@_d\,[\theta]\,\mathsf{m}$ correspond to in the lexicon. More specifically, what informational state semantics do the various types of prosodic markings contribute to the individual word categories. Since different pitch accent markings (θ and ρ) have different interpretations in a discourse context [Pierrehumbert and Hirschberg, 1990], let us have a brief look at the contribution of some of the more prominent prosodic markings we have dealt with in this work.

Table 5.1 summarizes the various pitch accent markings in Steedman's model. The second column of the table indicates the pitch accents tunes corresponding to these markings. Column three provides some examples of the phonological form in the lexicon.

Let us see the grammar implementation aspects of these markings.

5.3.1 The ρ-marking

Following the discussions of Pierrehumbert and Hirschberg [1990] and Steedman [2000a] in section 2.1 and section 2.2.1, we have learnt that a speaker's marking of an individual word with rheme tune renders it *salient* in the discourse context. Thereby, indicating the hearer(s) that the *open expression* is to be instantiated by the accented items and that the resulting proposition is to be *mutually believed* by the dialogue participants.

We propose that this contribution of rheme (ρ) tune to the contextual informational state of marked individual categories should be modeled in the grammar as part of the information structure dimension.

To represent this saliency and to-be-mutually believed informational state semantics of rheme tunes we employ the combination of feature-value pairs $\langle Rheme \rangle$true and $\langle Focus \rangle$true, and specify them in the semantic dimension of the word categories. Besides rendering the individual word salient in the discourse, the high rheme tune H* further indicates the speaker's commitment to the propositional content conveyed by the marked word. By commitment, we mean that the speaker is certain that the marked item is the one with which the open expression is to be instantiated, and to be added to the mutual beliefs by the dialogue participants. This further highlights that the speaker is uncontentious about the marked item. To represent the informational state of a marked item with respect to a speaker's *commitment* and *agreement* in the HLDS relational structure we use feature $\langle Owner \rangle$ and $\langle Agreed \rangle$ respectively.

Steedman [2000a]We use feature-value $\langle Owner \rangle$speaker to indicate speaker's commitment to the propositional content, and feature-value $\langle Owner \rangle$hearer to indicate speaker's lack of commitment. Lack of commitment on part of the speaker suggests that the hearers should own up to resolving or accommodation or clarification of the marked item. Next, to represent a speaker's uncertainty we use feature value $\langle Agreed \rangle$minus, and feature-value $\langle Agreed \rangle$plus to convey a speaker's uncontentiousness.

Using these four feature-value pairs the information structure dimension of a ρ-marked category with H* pitch accent is represented in our lexicon as follows:

$$(87) \quad @_d\,[\rho]\,\mathsf{m} \vdash @_{d:sort}\,(\mathbf{m} \wedge \langle Rheme \rangle \mathsf{true} \wedge \langle Focus \rangle \mathsf{true} \wedge \langle Agreed \rangle \mathsf{plus} \wedge \langle Owner \rangle \mathsf{speaker})$$

Using this representation of IS, the example in (88) illustrates the lexical entry for a H* marked noun type category 'box'.

(88) $\mathsf{BOX\_H*} \vdash \mathsf{n} : @_{m:sort}\ (\mathbf{box}\ \wedge \langle Rheme \rangle \mathsf{true}\ \wedge \langle Focus \rangle \mathsf{true}$
$\wedge \langle Agreed \rangle \mathsf{plus}\ \wedge \langle Owner \rangle \mathsf{speaker})$

In a similar manner the contribution of rheme pitch accent L* can be represented in the lexicon. Marking of a word with a low pitch accent L* indicates a speaker's *lack of commitment* to the propositional content conveyed by the marked item. In such a context, the marked item is, nevertheless, rendered *salient* but the speaker *doesn't* suggest that the hearer(s) instantiate the open proposition with it and adds it to the mutual beliefs. In fact, the lack of commitment on the speaker's part indicates that he himself is unable to make such an instantiation with the marked item. Thus it becomes the onus of the hearers to verify and confirm whether or not such an instantiation with the marked item is possible.

Following this, we specify the informational state of a L* marked item with feature-values $\langle Agreed \rangle$minus and $\langle Owner \rangle$hearer in the HLDS structure in the semantic dimension. Thus, the lexicon entry of a L* marked entity differs from one with a H* marked with regard to agreement and ownership feature-values. Example (89) illustrates the lexicon entry for a L* pitch accent marked noun type category 'box'.

(89) $\mathsf{BOX\_L*} \vdash \mathsf{n} : @_{m:sort}\ (\mathbf{box}\ \wedge \langle Rheme \rangle \mathsf{true}\ \wedge \langle Focus \rangle \mathsf{true}$
$\wedge \langle Agreed \rangle \mathsf{minus}\ \wedge \langle Owner \rangle \mathsf{hearer})$

Remember that in Chapter 4, section **??**, we employed these four feature-value pairs to specify the contextual informational status of discourse entities as the four dimensions of the information structure in a HLDS relational structure. By employing these four feature-value pairs again for the HLDS structures in the IS dimensions of the lexicon entries, we directly map the four dimensions of information structure onto the grammar itself.

5.3.2 The θ-marking

The marking of an individual word with theme pitch accent conveys that the accented item and not some alternative related item should be mutually believed. Theme pitch accents are employed by a speaker to convey the salience of some scale, linking the accented item to other items salient in a hearer's mutual beliefs. We model the contextual informational state contributed by theme pitch accents with feature-values $\langle Rheme \rangle$false and $\langle Focus \rangle$true.

In addition to this, a word marked with theme accent L+H* further indicates that the speaker assumes the informational content to be *mutually known* among the dialogue participants. Such a mutual awareness could be due to previous mention or the pragmatics of the situation. A L+H* pitch accent can therefore also indicate presupposition on speaker's part and thereby requiring the hearer to (owning to) accommodate or retrieve the marked item.

Since the information is assumed to be already agreed upon and mutually known, the speaker is also assumed to be uncontentious about the marked items.

Following this, we represent the *commitment* and *agreement* aspect of a theme (θ) tune L+H* marked category with feature-values $\langle Agreed \rangle$plus and $\langle Owner \rangle$hearer.

The lexicon entry in (90) corresponds to a θ-marked category bearing the pitch accent L+H*.

(90) $\mathsf{GREEN\_L{+}H^*} \vdash \mathsf{adj} : @_{\mathrm{r:color}}\ (\mathbf{green}\ \wedge \langle Rheme \rangle \mathsf{false}\ \wedge \langle Focus \rangle \mathsf{true}\ \wedge \langle Agreed \rangle \mathsf{plus}\ \wedge \langle Owner \rangle \mathsf{hearer})$

Table 5.2 summarizes the information structure feature values for the six pitch accents tunes. The feature value assignment is derived from the analysis of Pierrehumbert and Hirschberg [1990]. Since not all of the tunes have been extensively studied the ownership feature values for some of the tunes indicated with $^{\#}$ vary in certain contexts. We refer the reader to Pierrehumbert and Hirschberg [1990] for further discussions.

Pitch Accent	Rheme	Focus	Agreed	Owner
H*	true	true	plus	speaker
L*	true	true	minus	hearer
H*+L	true	true	plus	speaker$^{\#}$
H+L*	true	true	minus	speaker$^{\#}$
L+H*	false	true	plus	hearer
L*+H	false	true	minus	hearer

Table 5.2: Contribution of pitch accents tunes to IS feature-values.

Having elaborated upon the implementation of a multi-level sign with prosodic and information structure dimensions, we now come to the second question: *how do we model the constraints that govern the unifications at the prosody dimension*? More specifically, how do we model the derivational constraints of ι and ϕ*-marking* as per Steedman's model. What these two syntactic markings imply is as the following:

Prosodic Rule 1. *Constituents which are prosodically unmarked may freely combine with non-boundary constituents bearing prosodic information (i.e. θ/ρ-marking). As an unmarked category unifies with a marked category and results in a phrasal category, the informativity status of marked categories is spread over the unmarked phrasal constituents.*

Prosodic Rule 2. *Multiple pitch accents may occur in an intonational phrase. That is, categories bearing pitch accent markings of the same type (ρ or θ) may unify and result in identically marked larger categories. On the other hand, a ρ-marked category can't directly unify with a θ-marked category.*

Prosodic Rule 3. *A boundary must combine with at least one pitch accent to its left. Thus, the ι or ϕ markings which result in the formation of various types of intonational phrases can only take place with marked categories.*

Prosodic Rule 4. *A complete intonational phrase may combine with either another complete intonational phrase or an intermediate phrase. Thus, theme and rheme marked individual categories may unify only after they have become part of their respective theme and rheme intonational phrases.*

These rules basically specify the unification constraints at the prosodic dimension of the multi-level signs. The task of grammar implementation therefore requires one to model the derivational constraints at the prosodic level in addition to those acting at the syntactic level. Since derivations in the CCG are governed mainly by the compositional rules acting at the syntactic dimensions, it is the only possible level for specifying the additional constraints for prosodic derivation. To achieve the constrained prosodic derivations as postulated by these rules we exploit the *syntactic features* in the CCG lexicon.

syntactic features

As a first step, we follow Steedman's model and add a syntactic feature, namely `INFO`, to the syntactic category of the lexicon entries to indicate the informational state contributed by the pitch accent marking. The lexicon representation in (91) illustrates this modification.

(91) $\mathsf{phonology\_prosody} \vdash \mathsf{syntax}\ [\texttt{INFO}{=}\top] : \mathsf{semantics} \wedge \mathsf{information\ structure}$

The theme pitch accent tones mark an individual item as part of the theme information unit, whereas the rheme pitch accent tones mark the item to be part of the rheme information unit. We use feature-values `th` and `rh` for feature `INFO` to represent these information structure status contributed by the pitch accents tunes. Examples in (92)-(94) illustrate the values for feature `INFO` corresponding to the various types of pitch accent and nulltone markings.

(92) $$\frac{\mathsf{BOX\_H^*}}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=rh}] \\ @_m\mathbf{box} \wedge @_d\,[\rho]\,\mathsf{m}\end{array}}$$

(93) $$\frac{\mathsf{BOX\_L{+}H^*}}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=th}] \\ @_m\mathbf{box} \wedge @_d\,[\theta]\,\mathsf{m}\end{array}}$$

(94) $$\frac{\mathsf{box}}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=nt}] \\ @_m\mathbf{box} \wedge @_d\,[\eta]\,\mathsf{m}\end{array}}$$

Therefore the values corresponding to feature `INFO` indicate whether an individual word in the lexicon bears Steedman's θ,ρ or η-markings.

Next, in order to model the derivational constraints postulated by **Prosodic Rule 1** and **Prosodic Rule 2**, we define a *feature-value hierarchy* for the syntactic feature `INFO`. Figure 5.2 illustrate this hierarchy. The nodes in this tree indicates the range of values feature `INFO` may take. The leaf nodes correspond to the pitch accent tones presented in Table 5.1. The nodes `th` and `rh` have a *subsumption*

feature-value hierarchy

subsumption

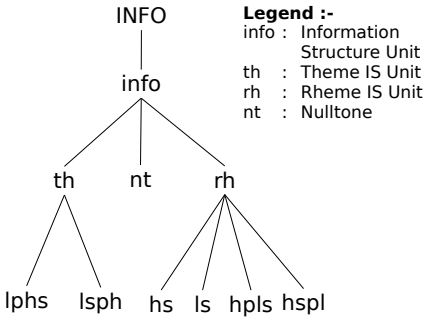


Figure 5.2: Feature-value hierarchy for syntactic feature `INFO`

relationship with the corresponding pitch accent tunes at the leaf nodes. Feature-value `th` subsumes the theme pitch accents values `lphs` and `lsph`, on the other hand feature-value `rh` subsumes the rheme pitch accents values `hs`, `ls`, `hpls` and `hspl`.

The relevance of this feature-value hierarchy for syntactic derivations is that: two marked word categories may unify only if a subsumption relationship holds between their `INFO` feature values. In this manner, a *θ-marked* category (`INFO=th`) may combine with another *θ-marked* category, but not with a *ρ-marked* category (`INFO=rh`). The same holds for *ρ-marked* categories.

However, we want to allow the words bearing nulltone i.e. those with feature-value `INFO=nt` for possible unifications with other *θ/ρ-marked* categories. For this reason, instead of specifying the feature-value for unmarked individuals as `INFO=nt`, which would block their unification with categories bearing `INFO=th` and `INFO=rh`, we choose to specify is as `INFO=`$\top$ in the lexicon. A $\top$ as feature value implies an underspecified value, and thus leaves the unmarked categories open for unification with other marked and unmarked categories.

With these additions to the syntactic category of the entries in the lexicon, the updated inventory of signs for the ongoing example derivation of robot response in (95) (repeated here from (69)), are specified as follows, in (96)-(99).

(95) R: (The BOX)$_{Th}$ (is GREEN)$_{Rh}$
L+H* LH% H* LL%

(96)
$$\frac{\text{The}}{\mathsf{s}_i\,[\texttt{INFO=}\top]\,/(\mathsf{s}_i\,[\texttt{INFO=}\top]\,\backslash\mathsf{np}_i\,[\texttt{INFO=}\top])/\mathsf{n}_t\,[\texttt{INFO=}\top]}$$
$$@_t\mathbf{the} \wedge @_t\langle \text{N}\rangle u \wedge @_d\,[\eta]\,\mathsf{t}$$

(97)
$$\frac{\text{BOX\_L+H*}}{\mathsf{n}_m\,[\texttt{INFO=th}]}$$
$$@_m\mathbf{box} \wedge @_d\,[\theta]\,\mathsf{m}$$

(98)
$$\frac{\text{is}}{\mathsf{s}_i\,[\texttt{INFO=}\top]\,\backslash!\mathsf{np}_x\,[\texttt{INFO=}\top]\,/\mathsf{adj}_y\,[\texttt{INFO=}\top]}$$
$$@_v\mathbf{is} \wedge @_v\langle \text{ACT}\rangle x \wedge @_v\langle \text{PAT}\rangle y \wedge @_d\,[\eta]\,\mathsf{v}$$

(99)
$$\frac{\text{GREEN\_H*}}{\begin{array}{c}\mathsf{adj}_r\,[\texttt{INFO=rh}]\\ @_r\mathbf{green}\wedge @_d\,[\rho]\,\mathsf{r}\end{array}}$$

With these modified entries in the lexicon, their combinations takes place as follows: the syntactic category for the sign in (96) seeks a category like (97) to its right. Additionally, the underspecified feature `INFO` in (96) permits it to unify with the *θ-marked* category for 'box' and result in a phrasal category, as shown in the derivation (100).

(100)
$$\dfrac{\dfrac{\text{The}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=}\top]\,/(\mathsf{s}_i\,[\texttt{INFO=}\top]\,\backslash\mathsf{np}_i\,[\texttt{INFO=}\top])/\mathsf{n}_t\,[\texttt{INFO=}\top]\\ @_t\mathbf{the}\wedge @_t\langle\text{N}\rangle u\wedge @_d\,[\eta]\,\mathsf{t}\end{array}} \quad \dfrac{\text{BOX\_L+H*}}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=th}]\\ @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\end{array}}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=th}]\,/(\mathsf{s}_i\,[\texttt{INFO=th}]\,\backslash!\mathsf{np}_i\,[\texttt{INFO=th}])\\ @_t\mathbf{the}\wedge @_t\langle\text{N}\rangle m\wedge @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\wedge @_d\,[\theta]\,\mathsf{t}\end{array}}>$$

Observe that in (100) the resulting phrasal category has the informational state of the marked category. This can be confirmed with the feature-value `INFO=th`, as well as the *θ-marked* structures in the semantic dimension.

In a similar manner, the lexical entries in (98) and (99) unify and result in another marked phrasal category. The derivation in (101) illustrates this unification.

(101)
$$\dfrac{\dfrac{\text{is}}{\begin{array}{c}\mathsf{s}_p\,[\texttt{INFO=}\top]\,\backslash!\mathsf{np}_x\,[\texttt{INFO=}\top]\,/\mathsf{adj}_y\,[\texttt{INFO=}\top]\\ @_v\mathbf{is}\wedge @_v\langle\text{ACT}\rangle x\wedge @_v\langle\text{PAT}\rangle y\wedge @_d\,[\eta]\,\mathsf{v}\end{array}} \quad \dfrac{\text{GREEN\_H*}}{\begin{array}{c}\mathsf{adj}_r\,[\texttt{INFO=rh}]\\ @_r\mathbf{green}\wedge @_d\,[\rho]\,\mathsf{r}\end{array}}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=rh}]\,\backslash!\mathsf{np}_x\,[\texttt{INFO=rh}]\\ @_v\mathbf{is}\wedge @_v\langle\text{ACT}\rangle x\wedge @_v\langle\text{PAT}\rangle r\wedge @_r\mathbf{green}\wedge @_d\,[\rho]\,\mathsf{r}\wedge @_d\,[\eta]\,\mathsf{v}\end{array}}>$$

Next, we observe that the resultant syntactic category of the signs in (100) seeks a category like (101) to its right. However, as no subsumption relationship holds between the values of the feature `INFO` in respective signs their unification is not permitted.

By using the syntactic feature `INFO` and defining a feature-value hierarchy we have thus modeled the constraints of **Prosodic Rule 1** and **Prosodic Rule 2**. As a consequence, individuals bearing no prosodic information may combine with non-boundary categories bearing rheme or theme pitch accent markings. Categories with ρ or θ-marking may combine with other categories bearing the same markings. However, two individual categories bearing ρ and θ respectively cannot unify. Unification of the theme and rheme marked phrasal categories is possible only when they have become part of respective intonational phrases as postulated by the **Prosodic Rule 4**.

5.3.3 The ι and ϕ-marking

What the ρ and θ-marked categories need in order to form intonational phrases is their combination with the *phrasal* and the *boundary tones* respectively. As

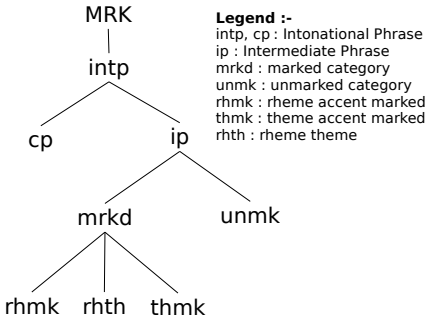


Figure 5.3: Feature-value hierarchy for syntactic feature `MRK`

discussed earlier, phrasal tones (L and H) and boundary tones (L% and H%) can be modeled as string elements in our lexicon. Their purpose is to only control the derivations, such as combining with marked categories to their left and result into intonational phrasal units.

Table 5.3 summarizes the inventory of phrasal and boundary tones in our lexicon. The syntactic category for the respective tones in the table follows from our earlier discussions that they delineate the intonational phrase boundaries.

Boundary Tones	**Lexicon**	**Syntactic Category**
Phrasal	L, H	$\mathsf{s}\$_\iota \backslash_\star \mathsf{s}\$_i$
Boundary	L%, H%	$\mathsf{s}\$_\phi \backslash_\star \mathsf{s}\$_\iota$
IpB	LL%, HL%, LH%, HH%	$\mathsf{s}\$_\phi \backslash_\star \mathsf{s}\$_i$

Table 5.3: Phrasal and Boundary tones

It is worth pointing out that from the combination of an intermediate phrase with a boundary tone that the boundary tones raise a phrasal tone to *intonational phrase boundaries*. For this reason we find it convenient to also have entries in our lexicon for intonational phrase boundaries (IpB) as such. An IpB is basically a boundary tone unified with a phrasal tone. The last row of Table 5.3 corresponds to IpB.

As postulated by **Prosodic Rule 3**, the boundary tones must combine with at least one pitch accent categories to their left. In order to model this constraint and derivational controls corresponding to ι and *ϕ-marking* in our grammar, we again follow the approach of using syntactic features. We introduce a syntactic feature namely `MRK` to indicate the type of marking on an individual category. The set of possible values for this feature corresponds to the ρ, θ, η, ι and ϕ-marking in Steedman's model. Figure 5.3 describes the feature-value hierarchy for feature `MRK`. The nodes in the tree indicate the values feature `MRK` can take.

Next, we specify the constraints of **Prosodic Rule 3** in the grammar by modifying the syntactic category of phrasal and boundaries with the syntactic feature `MRK` and `INFO`. Using the feature-value hierarchy, we constrain the type of phrasal

categories they can combine with to their left. The lexicon entry in (102), for example, illustrates the modified syntactic category for the phrasal tone L.

$$(102)\quad \frac{\mathsf{L}}{\mathsf{s}\,[\texttt{INFO=th,MRK=ip}]\,\$_\phi \backslash_\star \mathsf{s}\,[\texttt{INFO=th,MRK=thmk}]\,\$_i}$$

The feature-value `MRK=thmk` in the argument of the syntactic category in (102) suggests that the phrasal tone L may combine with a theme pitch accent marked category to its left (i.e. `INFO=th`). The feature-value `MRK=ip` in the domain suggests that the resultant category is an intermediate phrase, and it has the informativity of the theme marked category. In a similar manner, the lexicon entries for phrasal tones L taking a rheme marked category to its left can be defined.

The syntactic categories for non-boundary type entries in the lexicon are also updated for feature `MRK`. The lexical entry in (103) is a modified entry for an unmarked category. The feature-value `MRK=rhth` for unmarked categories *constrains* them from combining with the phrasal tones as there holds no subsumption relation among phrasal tone feature-values `MRK=thmk`, `MRK=rhmk` and unmarked category feature-value `MRK=rhth`.

$$(103)\quad \frac{\mathsf{The}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth}]\,/(\mathsf{s}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth}]\,\backslash\mathsf{np}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth}])/\\ \mathsf{n}_t\,[\texttt{INFO=}\top\texttt{,MRK=rhth}]\\ @_t\mathbf{the}\wedge @_t\langle \mathrm{N}\rangle u\wedge @_d\,[\eta]\,\mathsf{t}\end{array}}$$

A drawback of specifying unmarked categories with feature-value `MRK=rhth` is that they cannot unify with marked categories anymore. For example, the lexical entry in (103) will fail to unify with the theme pitch accent marked category in (104) containing `MRK=thmk`.

$$(104)\quad \frac{\mathsf{BOX\_L{+}H^*}}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=th,MRK=thmk}]\\ @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\end{array}}$$

As a solution to this problem, we resort to specifying the marked lexical entries with feature-value `MRK=mrkd`, instead of the tune specific values like `MRK=rhmk` and `MRK=thmk`. Since feature-value `MRK=mrkd` subsumes `MRK=rhth`, `MRK=rhmk` and `MRK=thmk`, marked categories may now combine with unmarked categories, as well as the phrasal tones. The following lexical entries illustrate these modifications.

$$(105)\quad \frac{\mathsf{BOX\_L{+}H^*}}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=th,MRK=mrkd}]\\ @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\end{array}}$$

$$(106)\quad \frac{\mathsf{GREEN\_H^*}}{\begin{array}{c}\mathsf{adj}_r\,[\texttt{INFO=rh,MRK=mrkd}]\\ @_r\mathbf{green}\wedge @_d\,[\rho]\,\mathsf{r}\end{array}}$$

Subsequently, an intermediate phrase may combine with a boundary tone to result in an intonational phrase. The lexical entry in (107), for example, is the boundary tone L%, which may combine with a rheme pitch accent marked intermediate phrase as indicated by the feature-values `INFO=rh, MRK=ip` in its argument category. The feature-values `INFO=info,MRK=cp` on the resultant category indicate that unification of L% leads to formation of complete intonational phrases.

$$\text{(107)}\quad \frac{\text{L\%}}{\mathtt{s}\,[\mathtt{INFO{=}info,MRK{=}cp}]\,\$_{\phi}\backslash_{\star}\mathtt{s}\,[\mathtt{INFO{=}rh,MRK{=}ip}]\,\$_{\iota}}$$

By specifying the feature-values `MRK=rhth` for unmarked categories, we were able to constrain their unification with phrasal tones. However, this choice is not enough to constrain their unification with boundary tones. This is because a subsumption relation holds between the feature-value `MRK=ip` for argument categories of boundary tones and the feature-value `MRK=rhth` (see Figure 5.3).

To model this constrain postulated by **Prosodic Rule 3**, we again resort to the use of syntactic feature. We introduce the syntactic feature `PHR` to the lexicon entries for boundary and non-boundary categories. The set of possible values feature `PHR` can take are `mkp` (for marked phrase), `ump` (for unmarked phrase) and `pbt` (for phrasal or boundary tone). Figure 5.4 illustrates the feature-value hierarchy for these values.

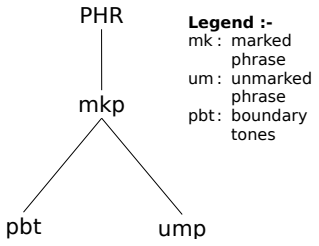


Figure 5.4: Feature-value hierarchy for syntactic feature `PHR`

To specify constraints on the derivations involving boundary tones, we use the feature-value hierarchy as follows. First we modify the lexical category for unmarked categories with feature-value `PHR=ump`. Lexical entry in (108) corresponds to the modified entry for (103).

$$\text{(108)}\quad \frac{\text{The}}{\begin{array}{c}\mathtt{s}_i\,[\mathtt{INFO{=}}\top\mathtt{,MRK{=}rhth,PHR{=}ump}]\,/(\mathtt{s}_i\,[\mathtt{INFO{=}}\top\mathtt{,MRK{=}rhth,PHR{=}ump}]\,\backslash\\ \mathtt{np}_i\,[\mathtt{INFO{=}}\top\mathtt{,MRK{=}rhth,PHR{=}ump}])/\mathtt{n}_t\,[\mathtt{INFO{=}}\top\mathtt{,MRK{=}rhth,PHR{=}ump}]\end{array}}{@_t\mathbf{the}\wedge @_t\langle \mathrm{N}\rangle u\wedge @_d\,[\eta]\,\mathtt{t}}$$

Next, we update the argument and resultant categories of the boundary tone with feature-value `PHR=pbt`. The lexical entry in (109) corresponds to these modification in the lexical entry (107) for boundary tone L%.

$$(109)\quad \frac{}{\mathsf{s}\,[\texttt{INFO=info,MRK=cp,PHR=mkp}]\,\$_{\phi}\backslash\!\!\star\mathsf{s}\,[\texttt{INFO=rh,MRK=ip,PHR=pbt}]\,\$_{\iota}}\;\mathsf{L\%}$$

Since no subsumption relation between feature-values `PHR=pbt` and `PHR=ump` holds, unmarked categories cannot unify with boundary tones anymore. However, to allow marked categories to unify with other unmarked categories and boundary tones, we modify the lexical entries for marked categories with feature value `PHR=mkp`, which subsumes `PHR=pbt` and `PHR=ump`. The modified lexical entries in (110) and (111) illustrate these modifications.

$$(110)\quad \frac{}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=th,MRK=mrkd,PHR=mkp}]\\ @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\end{array}}\;\mathsf{BOX\_L{+}H^*}$$

$$(111)\quad \frac{}{\begin{array}{c}\mathsf{adj}_r\,[\texttt{INFO=rh,MRK=mrkd,PHR=mkp}]\\ @_r\mathbf{green}\wedge @_d\,[\rho]\,\mathsf{r}\end{array}}\;\mathsf{GREEN\_H^*}$$

Before proceeding to the derivations involving phrasal and boundary tones, we need to answer another question, that is : *which of the phrasal and the boundary tones should unify with a marked phrasal category*? What rules in the grammar govern the decision choice that a marked phrasal category should combine with a low phrasal tone L or a high tone H? What rules specify that an intermediate phrase should combine with final-lowering boundary tones L% and not the high rising boundary tone H%?

Pierrehumbert and Hirschberg [1990] discusses the meaning contributed by phrasal tones L and H to the intermediate phrases (see section 2.1). Modeling these informational state contributions of phrasal tones in a grammar would require one to also model their discourse information structure semantics at a much finer level of discourse segments. However, in this thesis, we primarily focus at the level of full sentences, and therefore we are mainly concerned here with intonational phrase boundaries (IpBs) described in Table 5.3, namely LH%, LL% and HH%. The question, however, remains: *which of the intonational phrase boundary tones should unify with a marked phrasal categories*?

All boundary tones are *projections* of the rheme and theme information units on the phrase boundary, and delineate the intonational phrase boundaries [Steedman, 2000a]. Whether an IS unit is of rheme or theme type is indicated by the pitch accent types. Therefore, the IpB selection is eventually governed by the semantics of pitch accent bearing categories. But how exactly?

In section 5.2, while specifying the information status contribution of the pitch accent tunes on individual words, we used semantic features in the lexicon for representing speaker's *commitment* to the marked words. Using feature-value pairs $\langle Owner\rangle\mathsf{speaker}$ and $\langle Owner\rangle\mathsf{hearer}$ we represent whether it is the speaker or the hearer who commits itself to the instantiation of the open proposition with the marked item. It is these semantic features in the IS dimension of a marked category which govern the projection of boundary tones for respective IS units.

Steedman [2000a] argues that a L% boundary tone indicates a in the intonation contour on speaker's part, thereby expressing its desire to end the utterance and also marking the information conveyed by marked items as its contribution to the discourse. In such a case the *ownership* of (or the commitment to) the information lies with the speaker. On the other hand, a H% boundary tone indicates *forward-looking*, which suggests that the speaker intends to retain the hold and wants the hearer to pay attention to the information that follows. In such cases, the hearer is responsible for accommodating or retrieving (from mutual beliefs) the information conveyed so far. Therefore the ownership of (or the commitment to) the information resides with the hearer. For information or clarification questions, the final boundary tone H% besides indicating the hearer to retrieve the information, prompts him further for a response in return. The ownership of the information to be conveyed or clarified resides also on the hearer. To sum it up, a L% boundary indicates speaker's ownership of the information unit, whereas a H% boundary tone indicates the hearer's ownership.

final-lowering

Since it is the pitch accent tones which indicate the speaker's commitment to the marked item, the projection of boundary tone is also governed by them as follows: a speaker's assumption that some salient information content in the current discourse is mutually known to the hearer is conveyed by pitch accent L+H* in combination with intonational phrase boundary LH%. Similarly, a speaker's intention to question a piece of information is conveyed by pitch accent H* and IpB HH%, because it is the ownership of the hearer to resolve the accented items. On the other hand a speaker's intention to clarify uncertainty over a piece of information is conveyed by pitch accent L* in combination with IpB HH%. A speaker uses the pitch accent H* in combination with IpB LL% for asserting a piece of information.

Following this discussion it is not difficult to observe that in Steedman's IS theory the LL% and LH% boundary tones are respective projections of rheme and theme IS unit, which are owned by the speaker and hearer respectively. A *forward-looking* boundary tone HH% indicates the speaker's intention to seek a response from the hearer. Table 5.4 summarizes the relationship between pitch accents tunes, ownership and their projections as boundary tones.

Pitch Accent	**Ownership**	**Boundary Tones**
H*, H+L*, H*+L	Speaker	L%, LL%, HL%
L+H*, L*+H, L*	Hearer	H%, LH%, HH%

Table 5.4: Boundary tones as Ownership indicators of Pitch Accents tunes.

In order to model the constraints for derivations involving such combinations of boundary tones and marked categories, we again resort to the use of syntactic features. We introduce the `OWN` (for ownership) syntactic feature to the lexicon entries. The set of possible feature-values for feature `OWN` are `spkr` (for speaker) and `hear` (for hearer). Figure 5.5 illustrates the feature hierarchy for ownership. One can infer from the hierarchy that the speaker and hearer owned units cannot unify as such, just as rheme and theme marked phrases cannot unify directly.

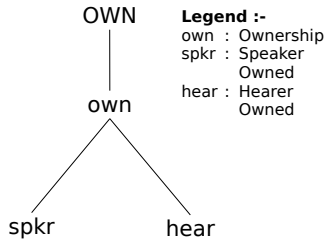


Figure 5.5: Feature-value hierarchy for syntactic feature OWN

With the introduction of syntactic features OWN, the lexical entries for the phrasal and boundary tones in Table 5.3 are revised as the following in (112)-(122).

(112)
$$\frac{}{\begin{array}{l}\mathtt{s}\,[\mathtt{INFO{=}rh,MRK{=}ip,PHR{=}mkp,OWN{=}}\top]\;\$_\iota \backslash_\star \\ \mathtt{s}\,[\mathtt{INFO{=}rh,MRK{=}rhmk,PHR{=}pbt,OWN{=}}\top]\;\$_i\end{array}}\;\mathsf{L}$$

(113)
$$\frac{}{\begin{array}{l}\mathtt{s}\,[\mathtt{INFO{=}th,MRK{=}ip,PHR{=}mkp,OWN{=}}\top]\;\$_\iota \backslash_\star \\ \mathtt{s}\,[\mathtt{INFO{=}th,MRK{=}thmk,PHR{=}pbt,OWN{=}}\top]\;\$_i\end{array}}\;\mathsf{L}$$

(114)
$$\frac{}{\begin{array}{l}\mathtt{s}\,[\mathtt{INFO{=}rh,MRK{=}ip,PHR{=}mkp,OWN{=}}\top]\;\$_\iota \backslash_\star \\ \mathtt{s}\,[\mathtt{INFO{=}rh,MRK{=}rhmk,PHR{=}pbt,OWN{=}}\top]\;\$_i\end{array}}\;\mathsf{H}$$

(115)
$$\frac{}{\begin{array}{l}\mathtt{s}\,[\mathtt{INFO{=}th,MRK{=}ip,PHR{=}mkp,OWN{=}}\top]\;\$_\iota \backslash_\star \\ \mathtt{s}\,[\mathtt{INFO{=}th,MRK{=}thmk,PHR{=}pbt,OWN{=}}\top]\;\$_i\end{array}}\;\mathsf{H}$$

(116)
$$\frac{}{\begin{array}{l}\mathtt{s}\,[\mathtt{INFO{=}info,MRK{=}cp,PHR{=}mkp,OWN{=}own}]\;\$_\iota \backslash_\star \\ \quad \mathtt{s}\,[\mathtt{INFO{=}rh,MRK{=}ip,PHR{=}pbt,OWN{=}spkr}]\;\$_i\end{array}}\;\mathsf{L\%}$$

(117)
$$\frac{}{\begin{array}{l}\mathtt{s}\,[\mathtt{INFO{=}info,MRK{=}cp,PHR{=}mkp,OWN{=}own}]\;\$_\iota \backslash_\star \\ \quad \mathtt{s}\,[\mathtt{INFO{=}th,MRK{=}ip,PHR{=}pbt,OWN{=}hear}]\;\$_i\end{array}}\;\mathsf{H\%}$$

(118)
$$\frac{}{\begin{array}{l}\mathtt{s}\,[\mathtt{INFO{=}info,MRK{=}cp,PHR{=}mkp,OWN{=}own}]\;\$_\iota \backslash_\star \\ \quad \mathtt{s}\,[\mathtt{INFO{=}rh,MRK{=}ip,PHR{=}pbt,OWN{=}hear}]\;\$_i\end{array}}\;\mathsf{H\%}$$

(119)
$$\frac{}{\begin{array}{l}\mathtt{s}\,[\mathtt{INFO{=}info,MRK{=}cp,PHR{=}mkp,OWN{=}own}]\;\$_\iota \backslash_\star \\ \mathtt{s}\,[\mathtt{INFO{=}rh,MRK{=}rhmk,PHR{=}pbt,OWN{=}spkr}]\;\$_i\end{array}}\;\mathsf{LL\%}$$

$$\frac{\text{LH\%}}{\mathsf{s}\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\;\$_\iota \backslash\!\star \quad \mathsf{s}\,[\texttt{INFO=th,MRK=thmk,PHR=pbt,OWN=hear}]\;\$_i} \tag{120}$$

$$\frac{\text{HH\%}}{\mathsf{s}\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\;\$_\iota \backslash\!\star \quad \mathsf{s}\,[\texttt{INFO=rh,MRK=rhmk,PHR=pbt,OWN=hear}]\;\$_i} \tag{121}$$

$$\frac{\text{HL\%}}{\mathsf{s}\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\;\$_\iota \backslash\!\star \quad \mathsf{s}\,[\texttt{INFO=rh,MRK=rhmk,PHR=pbt,OWN=spkr}]\;\$_i} \tag{122}$$

Similarly the lexicon entries for the non-boundary items in our ongoing example are also revised as the following (123)-(126). In order to avoid cluttering due to syntactic features across the categories in the syntactic dimension we use the shortned notation $[..]_\delta$, where the δ indicates indexing of features `INFO`, `MRK`, `PHR` and `OWN` in the domain of a category with those in the range: $[$`INFO=...,MRK=...,PHR=..., OWN=...`$]_\delta$.

$$\frac{\text{The}}{\mathsf{s}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/(\mathsf{s}_i\,[..]_\delta \backslash !\mathsf{np}_i\,[..]_\delta)/\mathsf{n}_t\,[..]_\delta \quad @_t\mathbf{the} \wedge @_t\langle \text{N}\rangle u \wedge @_d\,[\eta]\,\mathsf{t}} \tag{123}$$

$$\frac{\text{BOX\_L+H*}}{\mathsf{n}_m\,[\texttt{INFO=th,MRK=thmk,PHR=mkp,OWN=hear}] \quad @_m\mathbf{box} \wedge @_d\,[\theta]\,\mathsf{m}} \tag{124}$$

$$\frac{\text{is}}{\mathsf{s}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta \backslash !\mathsf{np}_x\,[..]_\delta\,/\mathsf{adj}_y\,[..]_\delta \quad @_v\mathbf{is} \wedge @_v\langle \text{ACT}\rangle x \wedge @_v\langle \text{PAT}\rangle y \wedge @_d\,[\eta]\,\mathsf{v}} \tag{125}$$

$$\frac{\text{GREEN\_H*}}{\mathsf{adj}_r\,[\texttt{INFO=rh,MRK=rhmk,PHR=mkp,OWN=spkr}] \quad @_r\mathbf{green} \wedge @_d\,[\rho]\,\mathsf{r}} \tag{126}$$

Now the combinations of these lexicon categories takes place as follows. The syntactic dimensions in (123) and (124) unify to results in the theme accent marked phrase "The BOX_L+H*" as follows:

$$\frac{\text{The}\quad\text{BOX\_L+H*}}{\mathsf{s}_i\,[\texttt{INFO=th,MRK=thmk,PHR=mkp,OWN=hear}]_\delta\,/(\mathsf{s}_i\,[..]_\delta \backslash !\mathsf{np}_i\,[..]_\delta) \quad @_t\mathbf{the} \wedge @_t\langle \text{N}\rangle m \wedge @_m\mathbf{box} \wedge @_d\,[\theta]\,\mathsf{m} \wedge @_d\,[\theta]\,\mathsf{t}}\,> \tag{127}$$

In a similar fashion, the categories in (125) and (126) unify to result in the rheme accent marked phrase "is GREEN_H*", as shown in derivation (128).

$$\frac{\text{is}\quad\text{GREEN\_H*}}{\mathsf{s}_i\,[\texttt{INFO=rh,MRK=rhmk,PHR=mkp,OWN=spkr}]_\delta \backslash !\mathsf{np}_x\,[..]_\delta \quad @_v\mathbf{is} \wedge @_v\langle \text{ACT}\rangle x \wedge @_v\langle \text{PAT}\rangle r \wedge @_r\mathbf{green} \wedge @_d\,[\rho]\,\mathsf{r} \wedge @_d\,[\rho]\,\mathsf{v}}\,> \tag{128}$$

Although the sign for phrasal category in (127) seeks a syntactic category such as (128) to its left, their unification is blocked due to that fact that their prosodic syntactic feature-values do not unify. Using syntactic features MRK and OWN, and a feature-value hierarchy for them we model the constraints postulated by **Prosodic Rule 3**.

The boundary tone (120) can now combine with the theme accent marked sign (127). The derivation in (129) illustrates the unification of the boundary tone with a pitch accent category to its left.

$$\frac{\text{The BOX\_L+H* LH\%}}{\begin{array}{c}\mathsf{s}_\phi\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]_\delta\,/(\mathsf{s}_\phi\,[..]_\delta\,\backslash!\mathsf{np}_i\,[..]_\delta) \\ @_t\mathbf{the}\wedge @_t\langle\text{N}\rangle m\wedge @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\wedge @_d\,[\theta]\,\mathsf{t}\end{array}}<\tag{129}$$

Observe the feature-values of the resultant phrase in (129). Feature-values INFO=info,MRK=cp,PHR=mkp,OWN=own, suggest that the phrase is now a complete intonational phrase, and can therefore unify with another complete intonational phrase.

Similarly, the rheme accent marked phrasal category in (128) unifies with the boundary tone (119) and result into another complete intonational phrase as follows:

$$\frac{\text{is GREEN\_H* LL\%}}{\begin{array}{c}\mathsf{s}_\phi\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]_\delta\,\backslash!\mathsf{np}_x\,[..]_\delta \\ @_v\mathbf{is}\wedge @_v\langle\text{ACT}\rangle x\wedge @_v\langle\text{PAT}\rangle r\wedge @_r\mathbf{green}\wedge @_d\,[\rho]\,\mathsf{r}\wedge @_d\,[\rho]\,\mathsf{v}\end{array}}>\tag{130}$$

Finally, the intonational phrase in (129) unifies with (130) and results in the complete sentence, which is again a complete intonational phrase. Derivation in (131) illustrates this unification.

$$\frac{\text{The BOX\_L+H* LH\%}\quad\text{is GREEN\_H* LL\%}}{\begin{array}{c}\mathsf{s}_\phi\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}] \\ @_v\mathbf{is}\wedge @_v\langle\text{ACT}\rangle t\wedge @_v\langle\text{PAT}\rangle r\wedge @_r\mathbf{green}\wedge @_t\mathbf{the}\wedge \\ @_t\langle\text{N}\rangle m\wedge @_m\mathbf{box}\wedge @_d\,[\rho]\,\mathsf{r}\wedge @_d\,[\theta]\,\mathsf{m}\wedge @_d\,[\theta]\,\mathsf{t}\wedge @_d\,[\rho]\,\mathsf{v}\end{array}}>\tag{131}$$

Observe the phonological level of the resultant category in (131). The surface form comprises word sequences and markings for prosodic information. Together they realize the underlying linguistic meaning and the information structure which has been compositionally built at the semantic dimension. The IS structures with θ-marking indicate the respective referent entities as belonging to the theme information unit in the current discourse. This is realized by the corresponding intonational tune in phrase "The BOX_L+H* LH%". On the other hand, the IS structures with ρ-marking suggest that the respective referents belong to the rheme information units. This is realized by the intonational tunes in phrase "is GREEN_H* LL%."

By introducing syntactic features INFO, MRK, PHR and OWN, and defining a feature-value hierarchy for them we have implemented the constraints of prosodic derivation as postulated in **Prosodic Rule 1**, **Prosodic Rule 2**, **Prosodic Rule 3** and **Prosodic Rule 4** in the OpenCCG platform. We have shown how the INFO

feature governs unifications of phrasal categories bearing rheme and theme informativity state. Using the OWN feature we are able to project the intonational phrase boundaries for rheme and theme intonational phrases. With feature MRK and PHR we have governed the construction of intonational phrases in a incremental manner, thereby enabling us to derive larger intonational phrases.

An interesting requirement of the combinatory prosody is the derivations involving prosodic bracketing which are *orthogonal* to the traditional surface derivations. In the following section, we describe our implementation for enabling the orthogonal bracketing sought by prosodic derivations.

5.4 Orthogonal Prosodic Bracketing

In order to elaborate on prosodic bracketing which is orthogonal to traditional surface derivations, we take the human-robot interaction in Figure 5.6 as an example. The dialogue turns for this interaction are shown in (132).

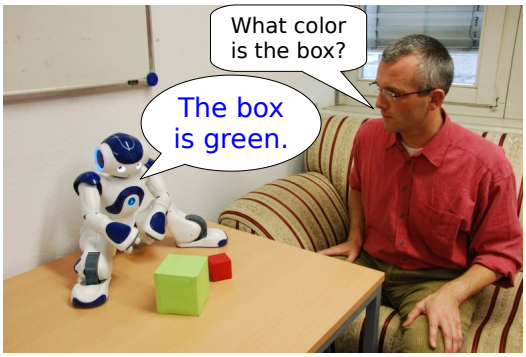


Figure 5.6: Human robot interaction corresponding to (132)

(132) a. H: What color is the box?

b. R: (The BOX is $)_{Th}$ (GREEN $)_{Rh}$
L+H* LH% H* LL%

It is noteworthy that the content of the robot utterance in (132) is exactly the same as that in (133) (from the scenario in (69)).

(133) a. H: Robot, what do you know about the box?

b. R: (The BOX $)_{Th}$ (is GREEN $)_{Rh}$
L+H* LH% H* LL%

These utterances, however, differ in their information structure partitioning, which is indicated by the bracketing. Therefore each of them establish a completely different contextual setup for their appropriateness. This can be inferred further from the respective human queries in (132a) and (133a). The question that we need

to address now is: *how do we model such prosodic derivations, as in (132b), which are orthogonal to the traditional surface derivations*?

The answer basically lies in the fact the both these utterances have different informational structure, and we need a different set of derivational rules for (132b) and (133b). As mentioned earlier, derivations in CCG are mainly governed by the combinatorics of syntactic dimension. In order to achieve the type of prosodic bracketing for (132b), with the expected surface form in (134) we need to see how a rheme intonational phrase such as "GREEN_H$*$ LL%" can be composed in our grammar.

(134) The BOX_L+H* is LH% GREEN_H* LL%

Observing the syntactic category of intonational boundary tone LL% in (119) we notice that it takes a syntactic category of type s as an argument. However, the syntactic category of modifiers like green, GREEN_H* or GREEN_L* in our grammar is a lexicon of type adj. So how do marked atomic categories (like nouns and adjectives) combine with boundary tones that seek marked phrasal categories of type s as argument?

type change rules

One way to go about it is introducing *type change rules* to the grammar for type-changing the syntactic category of atomic types into functional categories with a domain category of type s. Alternatively, we can simply introduce additional lexicon entries for atomic types with such functional categories. This is where the $ (dollar) notation used in syntactic categories for phrasal and boundary tones plays a vital role.

The type change rule in (135) is one such rule for type changing a marked atomic category of type adjective adj into a functional category s/!adj. Here, the modal operator ! in s/!adj ensures that category s/!adj doesn't unify with a adj on its right, that is, although functional the syntactic category s/!adj still behaves as an atomic category.

(135) *Type Change Rule*:
$\mathsf{adj}\$_1 \Rightarrow \mathsf{s/!adj}\$_1$

The rules allow us to type change atomic categories of type adj into a functional category s/!adj, as illustrated in (136).

(136)
$$\frac{\mathsf{GREEN\_H*}}{\begin{array}{c}\mathsf{adj}_r\ [\texttt{INFO=rh,MRK=mrkd}\\ \texttt{,PHR=mkp,OWN=spkr}]\\ @_r\mathbf{green} \wedge @_d\,[\rho]\,\mathsf{r}\end{array}} \Rightarrow \frac{\mathsf{GREEN\_H*}}{\begin{array}{c}\mathsf{s/!adj}_r\ [\texttt{INFO=rh,MRK=mrkd}\\ \texttt{,PHR=mkp,OWN=spkr}]\\ @_r\mathbf{green} \wedge @_d\,[\rho]\,\mathsf{r}\end{array}}$$

With such a type change rule in place, a marked atomic category may now unify with a boundary tone to its right and result, into a intonational phrase. In the current example the rheme boundary tone LL% (119) combines with the resultant category of (136) to its left and result into a rheme intonational phrase. The derivation in (137) illustrates this.

$$\frac{\text{GREEN\_H* LL\%}}{\mathsf{s}_\phi/!\mathsf{adj}_r\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\quad @_r\mathbf{green} \wedge @_d\,[\rho]\,\mathsf{r}}\,< \tag{137}$$

What we would want next is that this rheme intonational phrases unifies with the theme intonational phrase "The BOX_L+H* is LH%". Making this happen requires that the verbal head 'is' seeks a syntactic category "s/!adj" to its right. Given the lexicon entry in (138) for the verbal head 'is' which we have been using so far, we can't achieve this derivation. Therefore we introduce an alternate entry such as the one in (139) in our lexicon.

$$\frac{\text{is}}{\mathsf{s}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta \backslash !\mathsf{np}_x\,[..]_\delta\,/\mathsf{adj}_y\,[..]_\delta \quad @_v\mathbf{is} \wedge @_v\langle \text{ACT}\rangle x \wedge @_v\langle \text{PAT}\rangle y \wedge @_d\,[\eta]\,\mathsf{v}} \tag{138}$$

$$\frac{\text{is}}{\mathsf{s}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta \backslash !\mathsf{np}_x\,[..]_\delta\,/(\mathsf{s}_\phi/!\mathsf{adj}_y\,[..]_\delta) \quad @_v\mathbf{is} \wedge @_v\langle \text{ACT}\rangle x \wedge @_v\langle \text{PAT}\rangle y \wedge @_d\,[\eta]\,\mathsf{v}} \tag{139}$$

For every type change rule that we introduce in the grammar for enabling atomic categories to combine with boundary tones and form intonational phrases, we also need to introduce entries for verbal heads in our lexicon that seek these *ϕ-marked* categories.

Another noteworthy observation here is that the verbal head 'is' in robot utterance (132b) is part of the theme informational phrase "The BOX_L+H* is LH%", where as in (133b) it is part of the rheme information phrase "is GREEN_H* LL%". This distinction requires that the verbal head 'is' first unifies with its left complement, and forms a complete intonational phrase. Only then it may combine with its right complement. This is what makes the derivation in this sentence orthogonal to the traditional surface derivation.

Here again the syntactic feature `MRK` allows us to constrain the derivations. The lexicon entry in (139) cannot unify with the complete intonational phrase (137) anymore because the feature-value `MRK=cp` does not subsume the feature-value `MRK=rhth` for the unmarked verbal head (see Figure 5.3).

The derivation of the theme intonational phrase takes place as follows. The lexical entry (140) for the determiner 'The' unifies with (141) and result into the phrase "The BOX_L+H*", as illustrated in derivation (142)

$$\frac{\text{The}}{\mathsf{s}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/(\mathsf{s}_i\,[..]_\delta \backslash !\mathsf{np}_i\,[..]_\delta)/\mathsf{n}_t\,[..]_\delta \quad @_t\mathbf{the} \wedge @_t\langle \text{N}\rangle u \wedge @_d\,[\eta]\,\mathsf{t}} \tag{140}$$

$$\frac{\text{BOX\_L+H*}}{\mathsf{n}_m\,[\texttt{INFO=th,MRK=mrkd,PHR=mkp,OWN=hear}] \quad @_m\mathbf{box} \wedge @_d\,[\theta]\,\mathsf{m}} \tag{141}$$

$$\frac{\text{The}\quad \text{BOX\_L+H*}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=th,MRK=mrkd,PHR=mkp,OWN=hear}]_\delta\,/(\mathsf{s}_i\,[..]_\delta\,\backslash!\mathsf{np}_i\,[..]_\delta)\\ @_t\mathbf{the}\wedge @_t\langle\mathrm{N}\rangle m\wedge @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\wedge @_d\,[\theta]\,\mathsf{t}\end{array}}>\tag{142}$$

Next, the phrase "The BOX_L+H*" unifies with (139), the sign for verbal head 'is', and result in the phrase "The BOX_L+H* is", as shown in derivation (143).

$$\frac{\text{The}\quad \text{BOX\_L+H*}\quad \text{is}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=th,MRK=mrkd,PHR=mkp,OWN=hear}]_\delta\,/(\mathsf{s}_\phi/!\mathsf{adj}_y\,[..]_\delta)\\ @_v\mathbf{is}\wedge @_v\wedge\langle\mathrm{ACT}\rangle t\wedge @_v\langle\mathrm{PAT}\rangle y\wedge @_t\mathbf{the}\wedge\\ @_t\langle\mathrm{N}\rangle m\wedge @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\wedge @_d\,[\theta]\,\mathsf{t}\wedge @_d\,[\theta]\,\mathsf{v}\end{array}}>\tag{143}$$

The phrasal category in (143) has the informativity status of a theme unit, which can be seen from the *θ-marking* of the information structures in the semantic dimension. The theme marked phrasal category next, unifies with the intonational boundary tone LH% and results in a complete theme intonational phrase. The derivation in (144) illustrates this unification.

$$\frac{\text{The}\quad \text{BOX\_L+H*}\quad \text{is LH\%}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\,/\\ (\mathsf{s}_\phi/!\mathsf{adj}_y\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}])\\ @_v\mathbf{is}\wedge @_v\wedge\langle\mathrm{ACT}\rangle t\wedge @_v\langle\mathrm{PAT}\rangle y\wedge @_t\mathbf{the}\wedge\\ @_t\langle\mathrm{N}\rangle m\wedge @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\wedge @_d\,[\theta]\,\mathsf{t}\wedge @_d\,[\theta]\,\mathsf{v}\end{array}}>\tag{144}$$

The combination of boundary tone LH% with theme marked phrase (143) has the effect of modifying syntactic feature MRK=mrkd to MRK=cp. As a consequence, the resulting theme intonational phrase in (144) is now eligible for unification with the rheme intonational phrase in (137). Their unification results into another complete intonational phrase as following:

$$\frac{\text{The}\quad \text{BOX\_L+H*}\quad \text{is LH\%}\quad \text{GREEN\_H*}\quad \text{LL\%}}{\begin{array}{c}\mathsf{s}_\phi\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\\ @_v\mathbf{is}\wedge @_v\wedge\langle\mathrm{ACT}\rangle t\wedge @_v\langle\mathrm{PAT}\rangle r\wedge @_r\mathbf{green}\wedge\\ @_t\mathbf{the}\wedge @_t\langle\mathrm{N}\rangle m\wedge @_m\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{m}\wedge @_d\,[\theta]\,\mathsf{t}\wedge @_d\,[\theta]\,\mathsf{v}\wedge @_d\,[\rho]\,\mathsf{r}\end{array}}>\tag{145}$$

Observing the phonology and the semantic dimensions of the resultant category in (145), we see that the informational status of *θ-marked* entities $@_d\,[\theta]\,\mathsf{t}$, $@_d\,[\theta]\,\mathsf{m}$ and $@_d\,[\theta]\,\mathsf{v}$ is realized by the surface form "The BOX_L+H* is LH%". On the other hand, the information status of rheme information unit $@_d\,[\rho]\,\mathsf{r}$ is realized by surface from "GREEN_H* LL%".

The idea behind introducing type change rules for enabling atomic categories to unify with boundary tones, and making provisions for additional lexicon entries of verbal heads, which seek these type changed categories, is to allow prosodic derivations which are at times orthogonal to the traditional surface derivations. This can be seen in derivation (143) where the verb first unifies with its subject (142) and then with its object (137). Such unifications are not acceptable in traditional surface derivations, however, they are very important for information structure

based realization of prosody in surface forms. The multiple derivations engendered by CCG combinatorics under the constraints of the syntactic features we introduced allow us to achieve valid prosodic bracketing.

5.5 Examples Derivations

So far we have seen two different derivations for a robot response. In this section we illustrate derivations involving questions and clarification requests.

An Information Request

Consider the dialogue fragment in (146) corresponding to the interactive learning scenario in Figure 5.7. The lexicon entries in (147)-(152) enlist the lexicon for individual words participating in the derivation of the robot utterance in (146).

(146) R: What COLOR is the ball?
H* HH%
H: It is red.

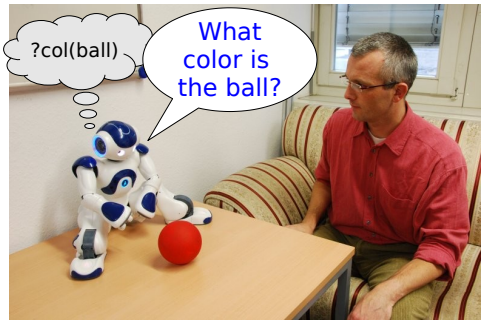


Figure 5.7: Seeking information.

(147)
$$\frac{\text{What}}{\mathsf{s}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/\Diamond(\mathsf{cop}_i\,[..]_\delta\,/\Diamond\mathsf{adj}_i\,[..]_\delta)/\mathsf{qclass}_q\,[..]_\delta \quad @_w\mathbf{what} \wedge @_w\langle\text{QAL}\rangle q \wedge @_d\,[\eta]\,\mathsf{w}}$$

(148)
$$\frac{\text{COLOR\_H*}}{\mathsf{qclass}_c\,[\texttt{INFO=rh,MRK=mrkd,PHR=mkp,OWN=hear}] \quad @_c\mathbf{color} \wedge @_d\,[\rho]\,\mathsf{c}}$$

(149)
$$\frac{\text{is}}{\mathsf{cop}_p\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/\mathsf{adj}_x\,[..]_\delta\,/\mathsf{np}_y\,[..]_\delta \quad @_v\mathbf{is} \wedge @_v\langle\text{ACT}\rangle y \wedge @_v\langle\text{PAT}\rangle x \wedge @_d\,[\eta]\,\mathsf{v}}$$

(150)
$$\frac{\text{the}}{\mathsf{np}_m\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/\Diamond\mathsf{n}_t\,[..]_\delta \quad @_t\mathbf{the} \wedge @_t\,\langle\text{N}\rangle\, b \wedge @_d\,[\eta]\,\mathsf{t}}$$

$$\frac{\text{ball}}{\begin{array}{c}\mathsf{n}_e\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top] \\ @_e\textbf{ball} \wedge @_d\,[\eta]\,\mathsf{e}\end{array}} \tag{151}$$

$$\frac{\text{HH\%}}{\begin{array}{c}\mathsf{s}\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\,\$_\iota\backslash\star \\ \mathsf{s}\,[\texttt{INFO=rh,MRK=rhmk,PHR=pbt,OWN=hear}]\,\$_i\end{array}} \tag{152}$$

The derivations in (153)-(157) illustrate the unification of these lexicon entries. First (147) unifies with (148) to result in (153).

$$\frac{\text{What COLOR\_H*}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=hs,MRK=mrkd,PHR=mkp,OWN=hear}]_\delta\,/\!\diamond(\mathsf{cop}_i\,[..]_\delta\,/\!\diamond\mathsf{adj}_i\,[..]_\delta) \\ @_w\textbf{what} \wedge @_w\langle\text{QAL}\rangle c \wedge @_c\mathsf{color} \wedge @_d\,[\rho]\,\mathsf{c} \wedge @_d\,[\rho]\,\mathsf{w}\end{array}}> \tag{153}$$

Next, signs in (150) and (151) unify as shown in (154):

$$\frac{\text{the ball}}{\begin{array}{c}\mathsf{np}_m\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top] \\ @_t\textbf{the} \wedge @_t\,\langle\text{N}\rangle\,e \wedge @_e\textbf{ball} \wedge @_d\,[\eta]\,\mathsf{e} \wedge @_d\,[\eta]\,\mathsf{t}\end{array}}> \tag{154}$$

Now sign (154) unifies with (149) to result in a verbal phrase as illustrated in (155):

$$\frac{\text{is the ball}}{\begin{array}{c}\mathsf{cop}_p\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/\mathsf{adj}_x\,[..]_\delta \\ @_v\textbf{is} \wedge @_v\langle\text{ACT}\rangle t \wedge @_v\langle\text{PAT}\rangle y \wedge @_t\textbf{the} \wedge @_t\,\langle\text{N}\rangle\,e \wedge @_e\textbf{ball} \\ \wedge @_d\,[\eta]\,\mathsf{e} \wedge @_d\,[\eta]\,\mathsf{t} \wedge @_d\,[\eta]\,\mathsf{v}\end{array}}> \tag{155}$$

Next, (155) combines with (153) to result in the full sentence, as shown in (156):

$$\frac{\text{What COLOR\_H*}\ \ \text{is the ball}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=rh,MRK=mrkd,PHR=mkp,OWN=hear}] \\ @_v\textbf{is} \wedge @_v\langle\text{ACT}\rangle t \wedge @_v\langle\text{PAT}\rangle w \wedge @_t\textbf{the} \wedge @_t\,\langle\text{N}\rangle\,e \wedge @_e\textbf{ball} \wedge @_w\textbf{what} \\ \wedge @_w\langle\text{QAL}\rangle c \wedge @_c\textbf{color} \wedge @_d\,[\rho]\,\mathsf{c} \wedge @_d\,[\rho]\,\mathsf{e} \wedge @_d\,[\rho]\,\mathsf{t} \wedge @_d\,[\rho]\,\mathsf{w} \wedge @_d\,[\rho]\,\mathsf{v}\end{array}}> \tag{156}$$

Next, (156) combines with boundary tone HH% and results in a complete intonational phrase.

$$\frac{\text{What COLOR\_H*}\ \ \text{is the ball HH\%}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}] \\ @_v\textbf{is} \wedge @_v\langle\text{ACT}\rangle t \wedge @_v\langle\text{PAT}\rangle w \wedge @_t\textbf{the} \wedge @_t\,\langle\text{N}\rangle\,e \wedge @_e\textbf{ball} \wedge @_w\textbf{what} \\ \wedge @_w\langle\text{QAL}\rangle c \wedge @_c\textbf{color} \wedge @_d\,[\rho]\,\mathsf{c} \wedge @_d\,[\rho]\,\mathsf{e} \wedge @_d\,[\rho]\,\mathsf{t} \wedge @_d\,[\rho]\,\mathsf{w} \wedge @_d\,[\rho]\,\mathsf{v}\end{array}}> \tag{157}$$

A Clarification Request

Consider the interactive learning scenario in Figure 5.8 corresponding to the dialogue fragment in (158). The entries in (159)-(164) enlist the lexicon for words participating in the derivation of the robot utterance in (158).

(158) R: Is that a RED ball?
L* HH%

H: No.

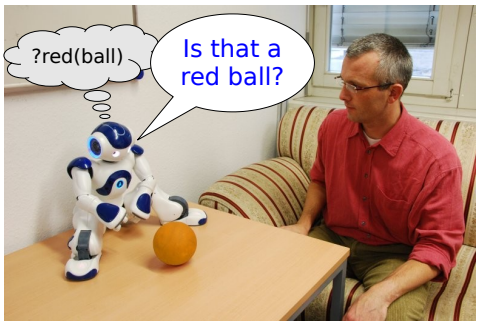


Figure 5.8: A clarification request.

(159)
$$\frac{\text{is}}{\mathsf{s}_p\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/\Diamond\mathsf{np}_x\,[..]_\delta\,/\Diamond\mathsf{np}_y\,[..]_\delta \quad @_v\mathbf{is}\wedge @_v\langle\text{ACT}\rangle y\wedge @_v\langle\text{PAT}\rangle x\wedge @_d\,[\eta]\,\mathsf{v}}$$

(160)
$$\frac{\text{that}}{\mathsf{np}_t\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top] \quad @_t\mathbf{that}\wedge @_d\,[\eta]\,\mathsf{t}}$$

(161)
$$\frac{\text{a}}{\mathsf{np}_m\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/\Diamond\mathsf{n}_t\,[..]_\delta \quad @_t\mathbf{a}\wedge @_t\,\langle\text{N}\rangle\,b\wedge @_d\,[\eta]\,\mathsf{t}}$$

(162)
$$\frac{\text{RED\_L*}}{\mathsf{n}_i\,[\texttt{INFO=rh,MRK=mrkd,PHR=mkp,OWN=hear}]\,/\mathsf{n}_r \quad @_r\mathbf{red}\wedge @_r\,\langle\text{OBJ}\rangle\,k\wedge @_d\,[\rho]\,\mathsf{r}}$$

(163)
$$\frac{\text{ball}}{\mathsf{n}_e\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top] \quad @_e\mathbf{ball}\wedge @_d\,[\eta]\,\mathsf{e}}$$

(164)
$$\frac{\text{HH\%}}{\mathsf{s}\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\,\$_\iota\backslash\star \quad \mathsf{s}\,[\texttt{INFO=rh,MRK=rhmk,PHR=pbt,OWN=hear}]\,\$_i}$$

The unifications in (165) to (169) illustrate the derivations for the utterance in (158).

(165)
$$\frac{\text{Is that}}{\mathsf{s}_p\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/\Diamond\mathsf{np}_x\,[..]_\delta \quad @_v\mathbf{is}\wedge @_v\langle\text{ACT}\rangle t\wedge @_v\langle\text{PAT}\rangle x\wedge @_t\mathbf{that}\wedge @_d\,[\eta]\,\mathsf{t}\wedge @_d\,[\eta]\,\mathsf{v}}\,>$$

(166)
$$\frac{\text{RED\_L* ball}}{\begin{array}{c}\mathsf{n}_i\,[\texttt{INFO=rh,MRK=mrkd,PHR=mkp,OWN=hear}]\\ @_r\mathbf{red} \wedge @_r\,\langle \text{OBJ}\rangle\, e \wedge @_e\mathbf{ball} \wedge @_d\,[\rho]\,\mathsf{e} \wedge @_d\,[\rho]\,\mathsf{r}\end{array}}>$$

(167)
$$\frac{\text{a RED\_L* ball}}{\begin{array}{c}\mathsf{np}_m\,[\texttt{INFO=rh,MRK=rhth,PHR=mkp,OWN=hear}]\\ @_t\mathsf{a} \wedge @_t\,\langle \text{N}\rangle\, r \wedge @_r\mathbf{red} \wedge @_r\,\langle \text{OBJ}\rangle\, e \wedge @_e\mathbf{box} \wedge @_d\,[\rho]\,\mathsf{e} \wedge @_d\,[\rho]\,\mathsf{r} \wedge @_d\,[\rho]\,\mathsf{t}\end{array}}>$$

(168)
$$\frac{\text{Is that a RED\_L* ball}}{\begin{array}{c}\mathsf{s}_p\,[\texttt{INFO=rh,MRK=mrkd,PHR=mkp,OWN=hear}]\\ @_v\mathbf{is} \wedge @_v\langle \text{ACT}\rangle t \wedge @_v\langle \text{PAT}\rangle f \wedge @_f\mathbf{a} \wedge @_t\,\langle \text{N}\rangle\, r \wedge @_r\mathbf{red} \wedge @_r\,\langle \text{OBJ}\rangle\, e\\ \wedge @_e\mathbf{box} \wedge @_t\mathbf{that} \wedge @_d\,[\rho]\,\mathsf{e} \wedge @_d\,[\rho]\,\mathsf{r} \wedge @_d\,[\rho]\,\mathsf{t} \wedge @_d\,[\rho]\,\mathsf{f} \wedge @_d\,[\rho]\,\mathsf{v}\end{array}}>$$

(169)
$$\frac{\text{Is that a RED\_L* ball HH\%}}{\begin{array}{c}\mathsf{s}_p\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\\ @_v\mathbf{is} \wedge @_v\langle \text{ACT}\rangle t \wedge @_v\langle \text{PAT}\rangle f \wedge @_f\mathbf{a} \wedge @_t\,\langle \text{N}\rangle\, r \wedge @_r\mathbf{red} \wedge @_r\,\langle \text{OBJ}\rangle\, e\\ \wedge @_e\mathbf{box} \wedge @_t\mathbf{that} \wedge @_d\,[\rho]\,\mathsf{e} \wedge @_d\,[\rho]\,\mathsf{r} \wedge @_d\,[\rho]\,\mathsf{t} \wedge @_d\,[\rho]\,\mathsf{f} \wedge @_d\,[\rho]\,\mathsf{v}\end{array}}>$$

Another Clarification Request

Consider the interactive scenario in Figure 5.9 where there are two objects present in the current scene. The dialogue fragment corresponding to it is shown in (170). The lexical entries from (171) to (176) participate in the derivation of the robot's clarification request in (170).

(170) H: This is a box.
R: (Is the BOX)$_{Th}$ (RED?)$_{Rh}$
L+H* LH% L* HH%
H: Yes.

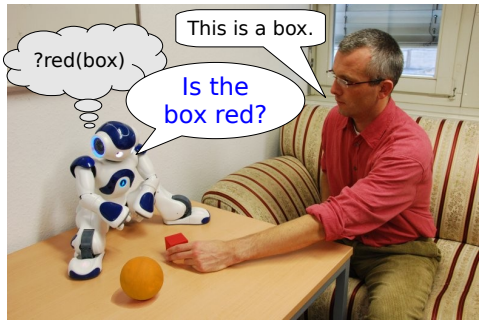


Figure 5.9: Clarifying color information.

(171)
$$\frac{\text{is}}{\begin{array}{c}\mathsf{s}_p\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/(\mathsf{s}/!\mathsf{adj}_y\,[..]_\delta)/\mathsf{np}_x\\ @_v\mathbf{is} \wedge @_v\langle \text{ACT}\rangle x \wedge @_v\langle \text{PAT}\rangle y \wedge @_d\,[\eta]\,\mathsf{v}\end{array}}$$

(172)
$$\frac{\text{the}}{\begin{array}{c}\mathsf{np}_m\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/\Diamond\mathsf{n}_t\,[..]_\delta\\ @_t\mathbf{the} \wedge @_t\,\langle \text{N}\rangle\, b \wedge @_d\,[\eta]\,\mathsf{t}\end{array}}$$

$$\frac{\text{BOX\_L+H*}}{\begin{array}{c}\mathsf{n}_e\,[\texttt{INFO=th,MRK=mrkd,PHR=mkp,OWN=hear}]\\ @_e\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{e}\end{array}} \tag{173}$$

$$\frac{\text{LH\%}}{\begin{array}{c}\mathsf{s}\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\,\$_\iota\backslash\star\\ \mathsf{s}\,[\texttt{INFO=th,MRK=thmk,PHR=pbt,OWN=hear}]\,\$_i\end{array}} \tag{174}$$

$$\frac{\text{RED\_L*}}{\begin{array}{c}\mathsf{s}_\phi/!\mathsf{adj}_r\,[\texttt{INFO=rh,MRK=mrkd,PHR=ump,OWN=hear}]\\ @_r\mathbf{red}\wedge @_d\,[\rho]\,\mathsf{r}\end{array}} \tag{175}$$

$$\frac{\text{HH\%}}{\begin{array}{c}\mathsf{s}\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\,\$_\iota\backslash\star\\ \mathsf{s}\,[\texttt{INFO=rh,MRK=rhmk,PHR=pbt,OWN=hear}]\,\$_i\end{array}} \tag{176}$$

The unification of these signs for the derivation of robot utterance in (170) takes place as follows. First the signs in (172) and (173) unify to result in noun phrase "the BOX_L+H*"

$$\frac{\text{the BOX\_L+H*}}{\begin{array}{c}\mathsf{np}_m\,[\texttt{INFO=th,MRK=mrkd,PHR=mkp,OWN=hear}]\\ @_t\mathbf{the}\wedge @_t\,\langle\text{N}\rangle\,e\wedge @_e\mathbf{box}\wedge @_d\,[\theta]\,\mathsf{e}\wedge @_d\,[\theta]\,\mathsf{t}\end{array}}> \tag{177}$$

Next, the sign in (177) combines with (171) and results in a verb phrase as illustrated in (178).

$$\frac{\text{is the BOX\_L+H*}}{\begin{array}{c}\mathsf{s}_p\,[\texttt{INFO=th,MRK=mrkd,PHR=mkp,OWN=hear}]_\delta\,/(\mathsf{s}/!\mathsf{adj}_y\,[..]_\delta)\\ @_v\mathbf{is}\wedge @_v\langle\text{ACT}\rangle t\wedge @_v\langle\text{PAT}\rangle y\wedge @_t\mathbf{the}\wedge @_t\,\langle\text{N}\rangle\,e\wedge @_e\mathbf{box}\wedge\\ @_d\,[\theta]\,\mathsf{e}\wedge @_d\,[\theta]\,\mathsf{t}\wedge @_d\,[\theta]\,\mathsf{v}\end{array}}> \tag{178}$$

The boundary tone in (174) unifies with (178) and results in a theme intonational phrase as follows:

$$\frac{\text{is the BOX\_L+H* LH\%}}{\begin{array}{c}\mathsf{s}_\phi\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]_\delta\,/(\mathsf{s}/!\mathsf{adj}_y\,[..]_\delta)\\ @_v\mathbf{is}\wedge @_v\langle\text{ACT}\rangle t\wedge @_v\langle\text{PAT}\rangle y\wedge @_t\mathbf{the}\wedge @_t\,\langle\text{N}\rangle\,e\wedge @_e\mathbf{box}\wedge\\ @_d\,[\theta]\,\mathsf{e}\wedge @_d\,[\theta]\,\mathsf{t}\wedge @_d\,[\theta]\,\mathsf{v}\end{array}}> \tag{179}$$

On the other side, the boundary tone in (176) combines with (175) to its left and results in rheme intonational phrase as illustrated in (180).

$$\frac{\text{RED\_L* HH\%}}{\begin{array}{c}\mathsf{s}_\phi/!\mathsf{adj}_r\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\\ @_r\mathbf{red}\wedge @_d\,[\rho]\,\mathsf{r}\end{array}}< \tag{180}$$

Finally the theme and rheme intonational phrase unify to result in the complete sentence.

$$\frac{\text{is the BOX\_L+H* LH\% RED\_L* HH\%}}{\begin{array}{c}\mathsf{s}_p\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}] \\ @_v\textbf{is} \wedge @_v\langle \textsc{act}\rangle t \wedge @_v\langle \textsc{pat}\rangle r \wedge @_t\textbf{the} \wedge @_t\,\langle \textsc{n}\rangle\, e \wedge @_e\textbf{box} \wedge @_r\textbf{red} \\ \wedge @_d\,[\rho]\,\mathsf{r} \wedge @_d\,[\theta]\,\mathsf{e} \wedge @_d\,[\theta]\,\mathsf{t} \wedge @_d\,[\theta]\,\mathsf{v}\end{array}}\;>$$

(181)

5.6 Limitations of Implementation

Although with our implementation of combinatory prosody in OpenCCG platform we have been able to achieve various prosodic construction, it does have some limitations. One of the limitation in the current state of the grammar implementation pertains to the prosodic derivations involving non-final rheme information units. In the following section we discuss this limitation and offer a possible solution to it.

5.6.1 Non-final Rheme Phrases

Observe that the rheme intonational phrase of the robot utterance in (182) has a phrasal tone as its intonational phrase boundary marker. On the other hand a rheme-final utterance, such as those in (132b) and (170) get a LL% boundary tone as intonational phrase boundary marker.

In terms of the theories of compositional intonation of Steedman [2000a]; Pierrehumbert and Hirschberg [1990] what the utterance (182) suggests is that a rheme marked intermediate phrase may combine with a complete theme intonational phrases. This is also postulated by **Prosodic Rule 4** of combinatory prosody.

(182) H: Which object is green?
R: (The BOX)$_{Rh}$ (is GREEN)$_{Th}$
H* L L+H* LH%

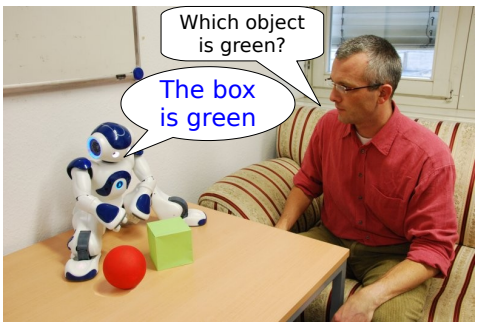


Figure 5.10: Non-final rheme units.

To enable such a derivation what we need in the grammar is a rule which raises a non-final rheme intermediate phrase to complete intonational phrase. We do not know how such a rule can be placed in the grammar. However, a possible solution could be as follows. It is at the level of information structure partitioning that the system can easily identify if a rheme IS unit is utterance final or non-final. When a rheme IS unit is non-final a semantic feature can be used to specify this in the information structure of an utterance.

Suppose we use a semantic feature $\langle NonFinal \rangle$ with value true to indicate non-finality of information unit. At the grammar side we map this semantic feature to a syntactic feature ORD with the set of possible values as nf (for non-final) and fl (for final). Next, we update the lexicon for rheme pitch accent marked categories with two entries for each, as in the following:

$$(183)\quad \frac{\textsf{BOX\_H*}}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=rh,MRK=mrkd,PHR=mkp,OWN=spkr,ORD=nf}]\\ @_m\mathbf{box} \wedge @_d\,[\rho]\,\mathsf{m}\end{array}}$$

$$(184)\quad \frac{\textsf{BOX\_H*}}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=rh,MRK=mrkd,PHR=mkp,OWN=spkr,ORD=fl}]\\ @_m\mathbf{box} \wedge @_d\,[\rho]\,\mathsf{m}\end{array}}$$

Next, we specify additional lexical entries for phrasal tones such that they would combine with non-final rheme marked categories and result into complete intonational phrases.

$$(185)\quad \frac{\textsf{L}}{\begin{array}{c}\mathsf{s}\,[\texttt{INFO=rh,MRK=cp,PHR=mkp,OWN=}\top\texttt{,ORD=nf}]\,\$_\iota \backslash_\star\\ \mathsf{s}\,[\texttt{INFO=info,MRK=mrkd,PHR=mkd,OWN=own,ORD=nf}]\,\$_i\end{array}}$$

Now, (185) may unify with (183) and result into a complete intonational phrase (MRK=cp) as shown in (186). The resultant phrase is now eligible for unification with complete theme intonational phrases.

$$(186)\quad \frac{\textsf{BOX\_H* L}}{\begin{array}{c}\mathsf{n}_m\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own,ORD=nf}]\\ @_m\mathbf{box} \wedge @_d\,[\rho]\,\mathsf{m}\end{array}}$$

5.6.2 Un-marked Theme Phrases

Another limitation in our grammar arises from those utterances where the theme information unit is unmarked. The example derivations that we have illustrated so far contained a marked theme, or were just rheme only utterances. With a marked theme and rheme IS partitioning, the realized intonational contour clearly indicates the intonational phrase partitioning. However, it is only appropriate to mark the theme with an L+H* pitch accent when it stands in contrast to a different established or accommodatable theme [Steedman, 2000a]. If the theme is unambiguously established in the context, it is common to find that it is deaccented throughout –as in the following exchange in (187) and its setup in Figure 5.11:

It is important to note that robot utterances in (187) and (188) are identical with respect to their information structure as far as the theme-rheme division goes. However, the context in Figure 5.12 with multiple salient objects necessitates a marked theme in (188), where as in Figure 5.11 the presence of only a single salient object doesn't necessitate a marked theme. We therefore need to distinguish the unmarked theme in the former from the unmarked theme in the latter.

(187) H: What color is the box?
R: (The box is)$_{Th}$ (GREEN)$_{Rh}$
H* LL%

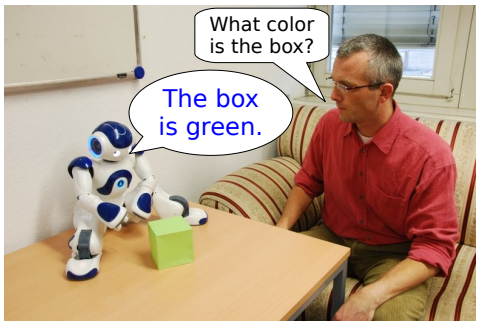


Figure 5.11: A single salient object.

(188) H: What color is the box?
R: (The BOX is)$_{Th}$ (GREEN)$_{Rh}$
L+H* LH% H* LL%

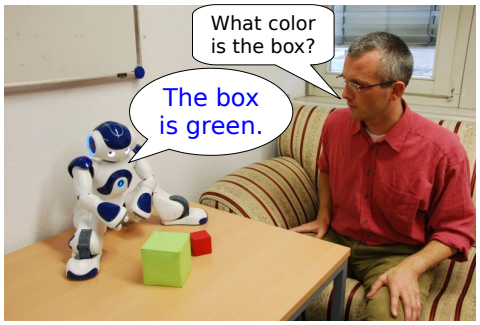


Figure 5.12: Multiple salient objects

Unmarked themes, unlike the marked variety, is always ambiguous with respect to information structure. From the intonation contour in (187) it is not obvious to infer the information structure partitioning. In the context of Figure 5.13 the same contour gives rise to an information structure in which "The box" is an unmarked theme, (189).

Given the grammar implementation we have we discussed so far we can obtain the following two constructs for derivation for unmarked theme utterance in (187).

(190)
$$\frac{\text{The box is}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta\,/(\mathsf{s}_\phi/!\mathsf{adj}_y\,[..]_\delta)\\ @_v\mathbf{is}\wedge @_v\wedge\langle\text{ACT}\rangle t\wedge @_v\langle\text{PAT}\rangle y\wedge @_t\mathbf{the}\wedge @_t\langle\text{N}\rangle m\wedge\\ @_m\mathbf{box}\wedge @_d\,[\eta]\,\mathsf{m}\wedge @_d\,[\eta]\,\mathsf{v}\end{array}}$$

(191)
$$\frac{\text{GREEN\_H* LL\%}}{\begin{array}{c}\mathsf{s}_\phi/!\mathsf{adj}_r\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=spkr}]\\ @_r\mathbf{green}\wedge @_d\,[\rho]\,\mathsf{r}\end{array}}$$

There are two limitations with these constructs. First, the informativity status of phrasal category in (190) is η (also INFO=$\top$). This is because we have modeled the

(189) H: What do you know about this box?
R: (The box)$_{Th}$ (is GREEN)$_{Rh}$
H* LL%

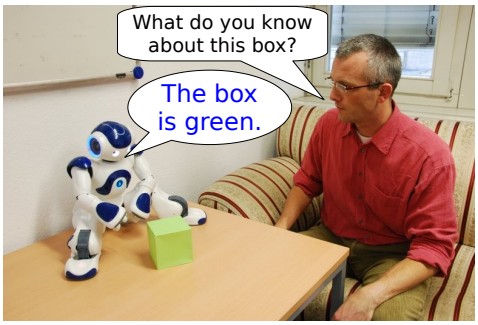


Figure 5.13: A single salient object

grammar such that marked categories contribute their *θ/ρ-marking* informativity to unmarked categories. This problem can be overcome by introducing two entries for each unmarked word with feature-values `INFO=rh` and `INFO=th` respectively. However, the feature-value `MRK=rhth` for unmarked categories will still constrain them from combining with complete intonational phrases. For this reason, the phrasal categories in (190) and (191) cannot unify.

What is basically needed here is a mechanism to treat unmarked phrasal categories as complete intonational phrases. This could be achieved by placing *type change rules* in the grammar to make unmarked phrasal categories complete intonational phrases.

(192) $\mathtt{s}\,[\mathtt{INFO}{=}\top,\mathtt{MRK}{=}\mathtt{rhth},\mathtt{PHR}{=}\mathtt{ump},\mathtt{OWN}{=}\top]\ \$ \Rightarrow$
$\mathtt{s}\,[\mathtt{INFO}{=}\mathtt{info},\mathtt{MRK}{=}\mathtt{cp},\mathtt{PHR}{=}\mathtt{mkp},\mathtt{OWN}{=}\mathtt{own}]\ \$$

Making such a type change rule would allow us to achieve the unification of (190) and (191). However, in the absence of sufficient grammar constraints, the type change rule may be triggered inadvertently to result into unwarranted derivations.

Another interesting case concerning the unmarked themes is that of questions as in the following:

(193) H: This is a box.
R: (What COLOR)$_{Rh}$ (is the box?)$_{Th}$
H* HH%

With our current implementation, the derivation for the robot utterance in (193) takes place as follows:

(194)
$$\frac{\text{What COLOR\_H*}}{\mathsf{s}_i\,[\mathtt{INFO}{=}\mathtt{rh},\mathtt{MRK}{=}\mathtt{mrkd},\mathtt{PHR}{=}\mathtt{mkp},\mathtt{OWN}{=}\mathtt{hear}]_\delta\,/\Diamond(\mathsf{cop}_i\,[..]_\delta\,/\Diamond\mathsf{adj}_i\,[..]_\delta)}$$
$$@_w\mathbf{what} \wedge @_w\langle \text{QAL}\rangle c \wedge @_c\mathbf{color} \wedge @_d\,[\rho]\,\mathsf{c} \wedge @_d\,[\rho]\,\mathsf{w}$$

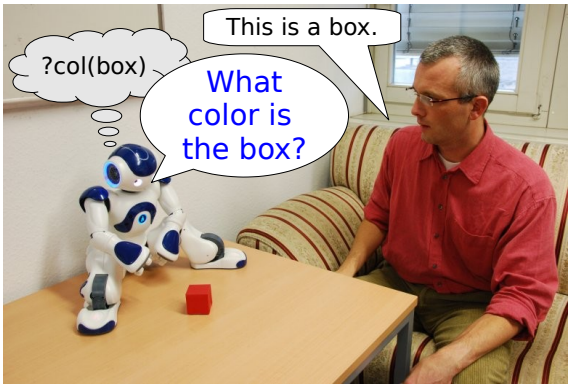


Figure 5.14: A single salient object

(195)
$$\frac{\text{HH\%}}{\begin{array}{c}\mathsf{s}\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]\;\$_\iota \backslash\!\!\star \\ \mathsf{s}\,[\texttt{INFO=rh,MRK=rhmk,PHR=pbt,OWN=hear}]\;\$_i\end{array}}$$

(196)
$$\frac{\text{What COLOR\_H* HH\%}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}]_\delta \,/\!\diamond (\mathsf{cop}_i\,[..]_\delta \,/\!\diamond \mathsf{adj}_i\,[..]_\delta) \\ @_w\textbf{what} \wedge @_w\langle \text{QAL}\rangle c \wedge @_e\textbf{color} \wedge @_d\,[\rho]\,\mathsf{c} \wedge @_d\,[\rho]\,\mathsf{w}\end{array}}<$$

(197)
$$\frac{\text{is the box}}{\begin{array}{c}\mathsf{cop}_p\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]_\delta \,/\mathsf{adj}_x\,[..]_\delta \\ @_v\textbf{is} \wedge @_v\langle \text{ACT}\rangle t \wedge \langle \text{PAT}\rangle y \wedge @_t\textbf{the} \wedge @_t\,\langle \text{N}\rangle\, e \wedge @_e\textbf{box} \\ \wedge @_d\,[\eta]\,\mathsf{e} \wedge @_d\,[\eta]\,\mathsf{t} \wedge @_v \wedge @_d\,[\eta]\,\mathsf{v}\end{array}}$$

Once again we observe that the phrasal category in (196) would rightly fail to unify with the unmarked phrasal category (197) to its right.

Employing a type change rule such as (198), as we did earlier, we could allow an unmarked theme category to unify with the intonational phrase to its left. The resultant category of this unification is illustrated in (199).

(198) $\mathsf{cop}\,[\texttt{INFO=}\top\texttt{,MRK=rhth,PHR=ump,OWN=}\top]\;\$ \Rightarrow$
$\mathsf{cop}\,[\texttt{INFO=info,MRK=cp,PHR=mkp,OWN=own}\,]\;\$$

(199)
$$\frac{\text{What COLOR\_H* HH\% is the box}}{\begin{array}{c}\mathsf{s}_i\,[\texttt{INFO=info,OWN=own,MRK=cp}] \\ @_v\textbf{is} \wedge @_v\langle \text{ACT}\rangle t \wedge @_v\langle \text{PAT}\rangle w \wedge @_t\textbf{the} \wedge @_t\,\langle \text{N}\rangle\, e \wedge @_e\textbf{box} \wedge \\ @_w\textbf{what} \wedge @_w\langle \text{QAL}\rangle c \wedge @_c\textbf{color} \wedge @_d\,[\rho]\,\mathsf{c} \wedge @_d\,[\rho]\,\mathsf{w} \wedge @_d\,[\rho]\,\mathsf{e} \wedge @_d\,[\rho]\,\mathsf{t}\end{array}}>$$

In this manner, although we achieve the unification of an unmarked theme and a complete intonational phrases, the intonational contour in (199) will fail to produce the intended result because the forward-reference boundary tone HH% is not at the end of the utterance.

What is additionally needed in a scenario like this is a mechanism to make the boundary tone of rheme-first-unmarked-theme questions constructions to get the intonational boundary tone of the rheme unit. In our implementation we do not yet have a mechanism for handling such requirements. We handle the prosodic generation for these cases by post-processing the surface form to result into intended intonation contours. Following that, the robot utterance (200a) is post processed to (200b).

(200) a. R: (What COLOR) (is the box)
H* HH%

b. R: (What COLOR) (is the box)
H* HH%

A similar issue arises with rheme-first-marked-theme information questions. As shown below, the rheme intonational phrase boundary are again post-processed.

(201) a. R: (What COLOR) (is the BOX)
H* HH% L+H* LH%

b. R: (What COLOR) (is the BOX)
H* L+H* HH%

We believe that a proper solution for handling the boundary tones for information questions lies with the semantic features such as $\langle SpeechAct \rangle$ or $\langle Mood \rangle$, which specify the communicative intention underlying the utterance. Modeling the equivalent of these post-processing rules in the grammar would require an interplay of these semantic features and the syntactic features we have introduced in our grammar.

5.7 Summary of the chapter

- In this chapter we discussed our approach to intonational realization of information structure. To achieve this we first introduced Steedman's model of combinatory prosody, and then discussed our approach to modeling it with multi-level signs in CCG framework. Then we elaborated our OpenCCG implementation of Steedman's θ, ρ, η, ι and *ϕ-markings*. To model them, we introduced syntactic features `INFO`, `MRK`, `PHR` and `OWN`, and defined a feature-value hierarchy for them.

  We have shown how feature `INFO` govern unifications of phrasal categories bearing rheme and theme informativity state. Using feature `OWN` we are able to project the intonational phrase boundaries for rheme and theme intonational phrases. With feature `MRK` and `PHR` we have governed the construction of intonational phrases in a incremental manner, thereby enabling us in deriving larger intonational phrases.

- We also introduced the grammar rules for achieving orthogonal prosodic bracketing in our implementation. Towards this, we introduce type change rules to enable atomic categories to combine with boundary tones. This also requires alternative lexical entries for verbal heads. Although this approach allows us to achieve the intended derivations, however, the alternative entries for verbal heads may become a derivation overhead as the grammar grows.

- Towards the end, we illustrated prosodic derivations for various types of utterances. We also disscussed the limitation of the current implementation, especially with unmarked themes. As discussed, most of these limitations result from the use of the OpenCCG platform. To cater to the needs of such prosodic derivations we need to explore how grammar rules can be effectively exploited.

Part III

Experimental Verification & Conclusions

6 Experimental Verification

In this chapter we describe our approach to experimental verification of the central claim made in this thesis. We start by briefly introducing various schemes for evaluating/measuring this work. Following this, we motivate the chosen methodology for the ongoing experiment. We elaborate on the experimental setup, the parameters, the design and the procedure. We conclude with a discussion on our findings and directions for future work.

6.1 Ascertaining the approach

The central claim of this thesis is that the contextual appropriateness of a robot's clarification requests, in a situated human-robot dialogue, has to also account for their contextually appropriate *intonation*. At the onset of this thesis, we have emphasized with various illustrations that contextually appropriate intonation enables a robot in presenting the intended meaning of its utterances. This provides for reducing the scope of ambiguities in a robot's utterances which may arise due to the *situatedness* of the dialogue.

The approach presented in this thesis follows from our claim that the contextually appropriate intonational realization of robot utterances can be established through the interplay of *intention* and *attention*, relative to a robot's *belief models*.

The approach developed in this thesis can be measured up along two lines of work. One line of work is to evaluate the developed system for ascertaining the aforementioned claim that contextually appropriate intonation in robot utterances enhances their contextual interpretability, and thereby reduces the scope of ambiguity in a situated dialogue. This would require the developed system to be subjected to third party testings.

In such a trial, a human user would be required to interact with a robot as part of one of the continuous interactive learning scenarios (see section 1.3). The user would be asked to perform the trials twice. Once with a system that doesn't explicitly model intonation in robot's clarification utterances i.e. relies on the default intonation produced by a text-to-speech synthesizer. Next, the user would have the same interaction with the same system running our approach to model contextually appropriate intonation. These two interactions can be used to draw various subjective and objective measures, e.g., which of the two instances of the

system did the subjects prefer, or interaction with which system were judged more coherent and efficient in meeting a learning tasks.

While such an evaluation will certainly bring to light the significance of intonation in robot's clarification requests, another line of work seems to offer more interesting avenues for getting insights into the role of *context* in establishing the appropriateness of an utterance's intonation. An experimental investigation in this direction will help us ascertain our approach to *information structure* assignment based on the contextual informational state of a robot's beliefs (cf. Chapter 4).

The cognitive system developed for George scenario, in year one of the CogX project (cf. section 1.3), is still primitive and can be operated and tested by its own developers. Therefore conducting the system evaluation as suggested earlier would not serve its purpose. Instead we pursue the second line of work i.e. experimental verification of the role of context in establishing the appropriateness of an utterance's intonation.

Earlier studies conducted in this direction include [Kruijff-Korbayová *et al.*, 2003], [White *et al.*, 2004b] and others. Both these studies were perception experiments and their findings confirm that utterances produced with contextually appropriate intonation are preferred more often than utterances with default intonation produced by a text-to-speech synthesizer in that context. These findings are relevant to our work because as with our system, both of these studies were conducted on systems (for question-answering) that produce intonation based on Steedman's [2000a] theory of IS.

However, where our approach differ from these two systems is that, firstly, their approaches use the preceding wh-question or dialogue to determine the context and for assigning information structure to a system response. The discourse context in these systems is therefore composed of only the dialogue history, however in our work the situatedness of the dialogue also makes the visual scene an inseparable part of the discourse context. Secondly, unlike with these system where the preceding dialogue or wh-question establishes the context for the hearer to interpret a system response, a robot's clarification requests in a situated dialogue doesn't necessarily have to relate to some preceding dialogue. For example, a clarification request may concern an unknown object that has been just introduced in the visual scene.

The task of producing contextually appropriate intonation in a robot's utterances has to therefore also account for the visual context, in addition to the dialogue context. As discussed in Chapter 4, our approach to IS assignment and intonation realization accounts for the visual as well as dialogue context. In the following section we outline the scheme of experimental studies that we want to conduct for verification of the approach presented in this thesis. Following this, we discuss our ongoing experimental study on investigating the role of the visual context in establishing the appropriateness of intonation.

6.2 Experimentation schemes

In order to verify the role of discourse context (visual and dialogue context) in establishing the appropriateness of intonation in utterances, we envisage the following dimensions along which we can test the intonation of a robot's clarifications utterances. Some of these ideas follow from the earlier studies in [Kruijff-Korbayová *et al.*, 2003] and [White *et al.*, 2004b].

Presence of a specific context An utterance can be evaluated in different contexts.

1. In a *neutral-context* scenario, system utterances are presented to the subjects without any particular visual or verbal context, i.e. "out of the blue". The judgements in a neutral-context will allow us to obtain a measure on the subjective qualitative judgments for utterances varying in intonation patterns. The scores for the various intonation patterns of an utterance will indicate the generally preferred intonation tune for an utterance. These scores can then be used as baseline for comparing the subjective qualitative judgement of an utterance in presence of specific context (visual or verbal, or visual and verbal both). Any reasonable improvement in the score of an utterance in the presence of specific context will allow us to infer that the particular intonation pattern for an utterance is more appropriate to the context.

2. In a *instant-context* scenario, one utterance with a certain visual context is presented to the subject at a time. The judgements in a instant-context will allow us to obtain a measure of subjective qualitative judgments for utterances in a exclusively visual context. The scores for the various intonation patterns of an utterance will indicate the generally preferred intonation tune for an utterance in a particular visual context. Any reasonable improvement in the score of an utterance in the instant-context with respect to the baseline score (from the neutral-context) allows us to infer that the particular intonation pattern is more appropriate to that visual context. On the other hand, a score lower than the baseline would suggest that the intonation is less appropriate in this visual scene.

3. In the *evolving-context* scenario, a sequences of utterances with visual context are presented, and thus the visual and verbal context evolves. The judgements in a evolving-context is basically about the last utterance. A comparative analysis of the measure of subjective qualitative judgments for utterances in a evolving-context with the baseline or instant-context will provide us insight into the influence of the interplay of visual and verbal context on the assignment of intonation.

Absolute vs. Relative judgement The subjective qualitative judgements of robot utterances can be *absolute* or *relative*. A judgement is absolute when a score

is assigned to an individual utterance, whereas they are relative when two or more utterance are compared, and a choice is made for the "better" alternative. Absolute judgments are easier to compare and group across the board, but there always is the problem that different subjects may have a differently set standard. One way to circumvent this problem is to take the difference between a judgment of an utterance with respect to a neutral context and a specific context. Relative judgments are more tangible and give more clearcut outcomes. However, the obtained judgments only pertain to the presented set of alternatives.

Active Involvement In the absence of an active involvement in the interaction with the system, subjects may suffer from being "detached", and may not react to nuances in system output realization in the same way as active interaction participants. Moreover, the "third-party" perspective precludes the use of any *objective* criteria, and therefore one can only elicit *subjective* judgments. One way to enable subjects to participate in the interaction is allow them to respond to the robot queries. This allows us to measure their reaction time, and the appropriateness/correctness of their responses.

Along these dimensions we have outlined a series of experiments for the purpose of the verification of our approach. In the following section we discuss the details of the ongoing experiment.

6.3 The Experiment

In this experiment, we seek to verify whether the visual context influences the placement of nuclear accent in an utterance. This is motivated from the established fact that nuclear accent in an utterance contrasts the marked individual from other competing alternatives (available due to their prior mention, or pragmatic accommodability) in a dialogue context [Pierrehumbert and Hirschberg, 1990], [Steedman, 2000a] (also see section 2.1.2 and 2.2). That is, whether or not the placement of nuclear accent in an utterance is appropriate and is governed by the competing alternatives available in the dialogue context.

Psycholinguistic studies of situated language processing have revealed that speakers look at objects shortly before mentioning them, while listeners tend to look at mentioned objects in their visual environment shortly after hearing the reference [Staudte and Crocker, 2009]. We hypothesize that in a situated dialogue, a listener's perception of the intonation tune of an utterance is also influenced by the content of the current visual scene.

For example, the clarification utterance, "Is the ball red?" when produced with the nuclear accent on the word 'ball' in a visual context where there is no other object in the common scene of the speaker and the hearer, is bound to trigger a sense of inappropriateness in the hearer's perception of the utterance. This is because placement of nuclear pitch on the word 'ball' in this sentence is only appropriate when some other (non ball type) object is present in the scene.

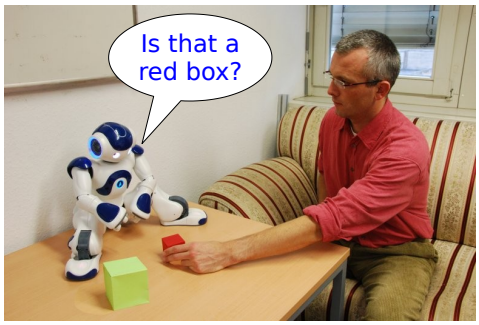


Figure 6.1: A *congruent* visual context.

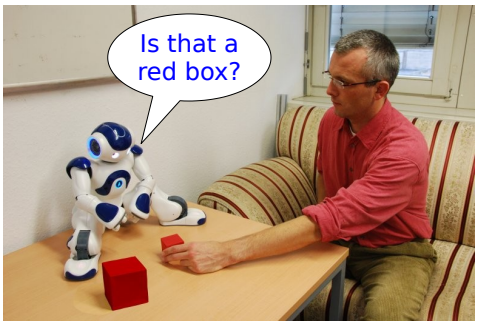


Figure 6.2: A *non-congruent* visual context.

The observations made above on the usage of contrast and placement of nuclear accent in a dialogue context also applies to a visual context where the presence of multiple objects in the visual scene of a robot, and hence the availability of competing visual properties govern the use of contrast and placement of nuclear accent in a robot's utterances.

Following this we argue that the placement of nuclear accent in robot utterance (202) is licenced in the visual context of Figure 6.1. The presence of a 'red' and a 'green box' prompts the use of contrast on the *color* property for distinguishing the intended box from the other.

(202) R: Is that a RED box?
L* HH%

On the other hand in the visual context of Figure 6.2, since both the boxes have the same color (i.e. red), the accent placement in (202) is not licenced and hence the usage is inappropriate. The nuclear accent placement in (203) utterance illustrates the appropriate nuclear accent placement in this visual context.

(203) R: Is that a red BOX?
L* HH%

In order to verify our claim, we have setup an experiment with the instant-context scenario scheme. A robot's clarification request corresponding to a visual scene is presented to subject for the judging the appropriateness of the utterance. The underlying hypothesis of the experiment is that:

Hypothesis 1. *If the comprehension is sensitive to the relationship of nuclear accent placement and the visual context than the variation in the placement of nuclear accent in an utterance can be perceived. The preferance of one pattern of accent placements over the other will provide us evidence in support of the role of visual context in establishing the appropriate intonation of an utterance.*

6.3.1 Methodology

Subjects

Thirty-one subjects, which included students and researchers, participated in the experiment. Out of these, twenty-one accessed the online version of the experiment, via the internet. The remaining ten undertook the experiment in our lab. Most of the participants were non-native speakers of English. Various psycholinguistic findings [Garbe *et al.*, 2003] reveal that the L2 speakers of English are equally sensitive to the intonational variations, though the tune interpretation varies with the individuals experience with the L2 language. On the basis of these findings we collapse data from both native and non-native English speakers. Participants were offered a sum of 5 Euros or an Amazon Gift Card worth 5 Euros for their successful completion of the experiment, provided they register. Additionally, three participants were drawn for the prize gift vouchers worth 20 Euros each.

Material and Design

Clarification request of the form "Is that a `color type`?" were chosen for the robot's clarification utterances, e.g. "Is that a red ball". The color and type values were selected such that they were monosyllabic words. This is done to maintain uniformity and avoid any other source of prosodic variation in the clarification request except for the contrastive placement. We used the following eight types (or shapes): *box*, *ball*, *ring*, *heart*, *disc*, *wedge*, *star* and *sphere*. Each of these shapes were made available in six colors: *red*, *blue*, *pink*, *green*, *brown* and *black*. Using these eight shapes and six colors, forty-eight clarification sentences in the aforementioned form were designed.

Each of the forty-eight sentences is then distributed over the three main experimental conditions of our experiment. The first condition captures the relationship between the visual context and the placement of nuclear accent in an utterance. Based on the presence or absence of multiple competitive properties in a scene the nuclear accent placement in an utterance was labeled as *congruent* (C) i.e. licenced by the visual scene, or *non-congruent* (NC) i.e. not licenced by the visual context. For example, the combination of accent placement in (202) and the visual scene Figure 6.1 correspond to a congruent experimental condition. On the other hand, the combination of accent placement in (203) and the visual scene Figure 6.1 correspond to a non-congruent condition. For inferring the role of visual context in acceptability of intonation tunes we hypothesis that:

Hypothesis 2. *If the comprehension is sensitive to the relationship of nuclear accent placement and the visual context then the utterances corresponding to the congruent conditions should be perceived more appropriate than utterances in a non-congruent condition.*

The second condition captures the placement of nuclear accent in an utterance. Two types of – *marked* and *unmarked*, nuclear accent placement were chosen. An

unmarked placement coincides with the assignment of nuclear accent to the last individual word in an utterance. This is also the default location of nuclear accent placement in a text-to-speech synthesizer. A marked nuclear accent placement doesn't correspond to this default position. We label the intonation contour resulting from a marked nuclear accent placement as tune B (as in (204a)) and the ones resulting from an unmarked nuclear accent placement as tune A (as in (204b)).

(204) a. R: Is that a RED box?
L* HH%

b. R: Is that a red BOX?
L* HH%

For inferring the role of visual context in acceptability of a *marked* vs. *unmarked* accent placement we hypothesis that:

Hypothesis 3. *If the comprehension is sensitive to the relationship of nuclear accent placement and the visual context then the marked and unmarked accent placements would be perceived more appropriate in congruent visual scenes than a non-congruent scenes.*

The third experimental condition correspond to whether the robot's hypothesis about the objects in the visual scene in the correct or incorrect. A robot's hypothesis indicates the beliefs it currently holds about the visual scene. Since a robot's perceptory senses are not perfect, its beliefs may or may not be the same as the human. In such an eventuality the human responses 'YES' or 'NO', indicate to the robot whether its perception about the world is correct or not. Another reason for introducing correct and incorrect robot hypothesis is to avoid bias in subject's judgement due to rightness or wrongness of the robot's clarifications. We hypothesis that:

Hypothesis 4. *If the comprehension is sensitive to the relationship of nuclear accent placement and the visual context then a subject's perception of the appropriateness of an utterance would not be affected by the correctness of the robot's hypothesis. That is, congruent and non-congruent stimuli would have the same score distribution for both 'YES' and 'NO' human responses.*

These three experimental conditions thus provide for 2x2x2=8 combinations of conditions in total. We represent these combinations as C-A-YES, C-A-NO, C-B-YES, C-B-NO, NC-A-YES, NC-A-NO, NC-B-YES and NC-B-NO. Each of the forty-eight sentences mentioned above are distributed over these conditions. This results into a stimuli comprising of 384 clarification requests.

Besides these eight experimental conditions we introduce two "filler" nuclear accent placements in the utterances. This is done to overcome the auditory saturation due to tune A and tune B in the stimuli. These filler tunes correspond to the accent placement on the referential expression "that" and the verbal head "is". We label them as tune D (as in (205a)) and tune C (as in (205b)) respectively.

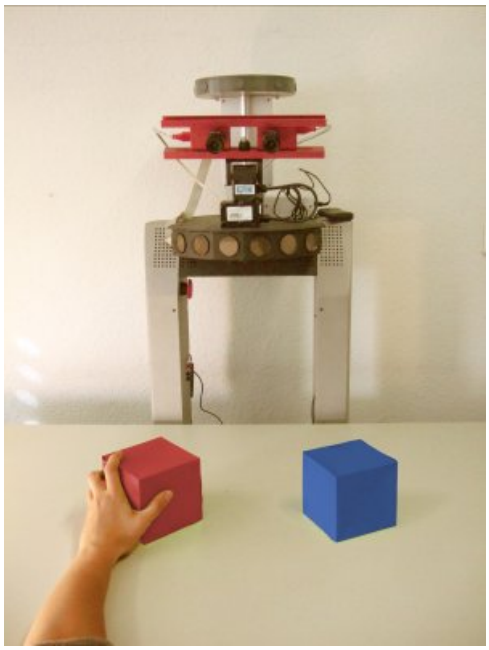


Figure 6.3: A *congruent* visual context for *marked* accent placement in (204a).

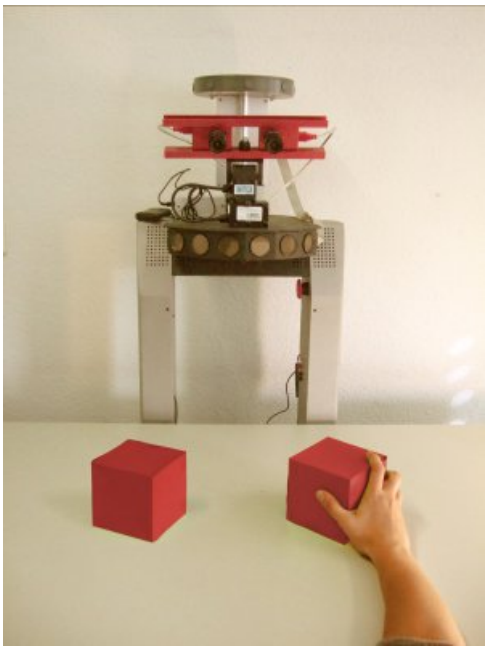


Figure 6.4: A *non-congruent* visual context for *marked* accent placement in (204a).

(205) a. R: Is THAT a red box?
L* HH%

b. R: Is that a red box?
L* HH%

The introduction of equally many filler tunes (i.e. 384 tune D and tune C in total) results into 768 utterances in the stimuli. The stimuli is then divided into eight different lists of ninety-six such that each list has all the forty-eight sentences, and evenly distributed over the eight conditions and two filler tunes.

For the presentation of visual context in the scenes, two (not necessarily different) shapes were paired in a picture (300x400 pixels), with a PeopleBot[1] standing at the table, see Figure 6.3 and 6.4. The pairing is done such that each of these shape occur as an object that is already present on the table, and as an object that is being introduced. Not all the combination of shapes were used as that would have resulted into a huge space of visual scenes. The sixteen pairs (where the first object represents the shape being introduced and the second object the shape that is already present) we have used are: *ball-ball*, *ball-heart*, *heart-ball*, *heart-heart*, *disc-disc*, *disc-cube*, *cube-disc*, *cube-cube*, *ring-ring*, *ring-star*, *star-ring*, *star-star*, *wedge-wedge*, *wedge-sphere*, *sphere-wedge*, and *sphere-sphere*.

[1]One of the robots for the George scenario in the CogX project.

The color of these object pairs were also chosen deliberately. We did not opt for all combination of colors pairs as that would again result into a huge space of visual scenes. The six color pairs (where the first color represent the color of the object being introduced and the second represent the color of the shape already present) we use are: *black-black*, *black-pink*, *pink-black*, *pink-pink*, *blue-blue*, *blue-red*, *red-blue*, *red-red*, *green-green*, *green-brown*, *brown-green*, *brown-brown*.

In addition to this, a left-hand and a right-hand orientation for the two objects on the table were used to avoid visual saturation or eye fixation.

These twelve color pairs and the two (left and right) orientations of the visual scene result into 12x2=24 visual scenes for each of the sixteen object pairs. This makes a total of 16x24=384 visual scenes, which is same as the size of stimuli (excluding the filler tunes).

Next, the stimuli of 768 utterances is divided into eight separate lists of ninety-six stimuli each. The rational behind this is that since we didn't use all possible object and color combinations in the visual scenes, we have 384 visual scenes for 768 utterances. This implies that each scene should be repeated no more than twice. However, by distributing the stimuli across eight lists we ensure that a subject never sees the particular shape and color object more than once for each of the eight conditions. For example, an utterance such as "Is that a red ball" is first distributed over the eight conditions, and each of these conditions is then placed in one of the eight stimuli lists. Such a distribution allows us to ensure that a subject can not assume or guess a particular shape and object combinations occurred with a particular condition in an experiment. The presentation of the conditions in each of these eight list is also distributed such that the subject cannot guess the next condition.

Audio files for 768 utterances were recorded using the MARY[2] text-to-speech synthesizer (TTS) [Schröder and Trouvain, 2003]. The MBROLA[3] 'mborla-us2' voice of a US-English male speaker were used for synthesizing robot's clarification requests. The TTS was indicated about the nuclear accent placement through the MARYXML format.

For the convenience of setting up the experiment, audio file for the human response with a 'YES' and 'NO' response were also synthesized using MARY, albeit with 'english-slt-arctic' voice (US-English female speaker SLT voice).

Procedure

A series of human-robot interactions were built for each of the eight stimuli lists using the combinations of these visual scenes and corresponding audio files. Following the design requirements of our experiment we have setup the experiment as a Web-Experiment. The experiment has been designed using the WebExp[4] system for conducting psychological experiments over the World Wide Web. The WebExp

[2] `mary.dfki.de`

[3] `http://tcts.fpms.ac.be/synthesis/`

[4] http://www.webexp.info/

server has been hosted on a server running Linux version 2.6.26-2-amd64 and having 1GB RAM. The Web-Experiment also offers us a possibility to reach native speakers of English for our experiment.

On arrival of subjects at the Web-Experiment web page, they were asked to read the instructions on the web page to familiarize themselves with the experiment. They were informed that they will see a series of pictures of a scene with a robot standing at a table with some objects on it. They will hear the robot asking a question and a human answering it. Every time they will make a judgment about the appropriateness of robot's question. Between robot scenes they will check the correctness of simple calculations.

Further instructions were provided on the content of a visual scene and the human-robot interaction. Subject were informed that in each robot scene there is one object already on the table that the robot knows about, and then another object is being presented by the human. The robot asks a question to verify whether it recognized correctly the type and the color of the object being shown. Since its recognition capacity is imperfect, it may make a mistake. The human responds to the robot with a 'YES' or a 'NO'. Subjects were asked to evaluate whether the robot asked the question in a way appropriate to the current scene, irrespective of whether it recognized the object (its type and color) correctly or not.

Participants were also instructed to close all heavy processing on their machines and network transactions for smooth conduction of the experiment. Additionally participants were asked to set the audio and visual setting of their machines to optimal level.

At the onset of the experiment, subject were asked for details regarding their age, gender, mother tongue, English they speak (US, UK, etc.), educational background, and their past experience with spoken language interfaces. After this a subject is automatically assigned one of the eight lists of stimuli. Next, the subject is introduced with the presentation style of a stimuli, the tasks the subject needs to perform, through a set of six practice stimuli. The stimuli for practice is evenly distributed over the eight conditions, only two shapes from the experiment, and no color from the experiment were used. Again this was done to avoid any bias for shapes, colors or conditions in the main experiment. Subjects were given the instruction once before the training and once again before the start of the main experiment.

In the practice session and the main experiment, the presentation of stimuli and the evaluation of the stimuli takes place in three steps.

First step, the visual stimuli (a picture) is presented to the subject, and with a delay of 1500ms the corresponding audio stimuli for the robot's clarification request is played. This added delay is a standard procedure for visual preview as visual stimuli capture a subject's visual attention. In the absence of a visual preview, linking the attention captured by the visual scene with the audio stimulus from the clarification would have been a challenging task for the subject. The sentence would be over before the participants would have started to pay attention to the spoken stimuli. Once the audio stops playing, the visual scene disappears after a delay of

1s. This delay is added to give the subject some time for linking the dialogue with the visual scene.

In the second step, the subject is asked to evaluate how appropriately the question was asked by the robot. The subject indicate their judgement by selecting a radio-button on a 5-point scale between good and bad. In the third step, the subjects were shown a simple math calculation task and were asked to judge whether or not the calculation was correct. An audio with a ticking of a clock was also played until the subject responded. The purpose of the calculation task and the clock audio is to break the subject's visual and audio stimulation, due to the current presentation, before proceeding to the next presentation. Once the subject responds to the calculation task, the next stimuli is presented as just described.

After the presentation of forty-eight stimuli, participants were instructed that they were half-way through the experiment and can have a short break of 1-2 minutes if they wish. Towards the end, the subject were able to register their email address for the gift vouchers and the prize draw. Subject were also able to provide additional feedback or queries on the experiment. The experiment was designed to take 20-25 minutes to finish.

To reach out to as many participants, we also ran the experiment in a on-site fashion. Interested participants were invited to our lab and were provided access to the Web-Experiment through a laptop, set up in one of the rooms. No additional instructions were given to these participants. They were simply directed to the Web-Experiment web-page and were asked to follow the instructions mentioned there.

6.3.2 Results

The results discussed here is a preliminary analysis of the ongoing experiment. The data collected here is from the thirty-one participants (i.e. 31x96=2976 data points for analysis). In this round we exclude the filler tones from the current analysis and investigate mainly into the eight experimental conditions (this makes 2976/2=1488 data points under analysis).

First we looked at the distribution of scores over all the conditions. This was done to get a first hand feel of the data. The score for 5 (704 data points) and 1 (162 data points) offered a very clear distribution for the third condition, that stimuli with 'YES' were preferred more often than stimuli with 'NO'. This findings contradict **Hypothesis 4**. However, these scores did not offer any insight into the other two conditions. The distribution of score for 4 (227 points) and 2 (173 data points) offered some interesting patterns indicating that the other two conditions play some role. The score for 3 (222 data points) is again unable to offer any clear direction. Following these observations we clubbed the score 5 and 4 under the label 'GOOD', whereas, the score of 1 and 2 were clubbed as 'BAD'. We assume the score of 3 indicates a subject's neutrality or undecidability and were hence labeled as 'NUTRL'. For the preliminary analysis we are mainly interested in the distribution of GOOD and BAD. We leave NUTRL for later work.

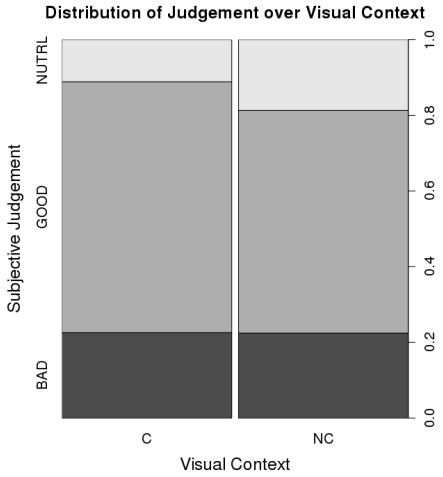


Figure 6.5: Distribution of Subjective Judgement over Visual Context.

In order to verify **Hypothesis 2** we investigated into the distribution of the GOOD and BAD judgments over all the congruent (C) and non-congruent (NC). Figure 6.5 provides a plot of the distribution of the subjective judgement of the stimuli over the congruent and non-congruent visual scenes. It can be inferred from these plots that utterances in a congruent visual context were more often judged good (1041 times) than bad (297 times). However, the distribution of judgement for the non-congruent visual context is not very different from the congruent context. Almost 60% of the stimuli in the non-congruent visual context were judged good (899 times). That is, although the pitch accent placement was not licenced by the visual context of the scenes, the utterances were often judged good. This is contrary to **Hypothesis 2**. We expected the subjective judgement of utterances in non-congruent visual contexts to be mostly bad. Although the figures for congruent and non-congruent doesn't differ by a huge margin they seem to provide some evidence for our hypothesis.

We investigated further to verify **Hypothesis 3**. The plot in Figure 6.6 provides the distribution of the subjective judgement over the marked and unmarked tunes B and A respectively. While the distribution suggests that both tunes A and B were equally preferred, the distribution of scores of tunes over the congruent and non-congruent should provide further evidence. As predicted tune A were judged GOOD 241 times in a congruent condition(C) than 223 times in a non-congruent condition(NC). Similarly tune B was judged GOOD 252 times in a congruent condition (C), than 215 times in the non-congruent condition.

Another way to verify **Hypothesis 3** is to look at the distribution of BAD score. Tune A has been judged 77 times BAD in congruent condition than 82 times in non-congruent condition i.e. more often scored BAD in non-congruent. This is as expected. Tune B on the other hand has been judged more often (91 times) BAD in congruent condition than the non-congruent condition (85 times). This is

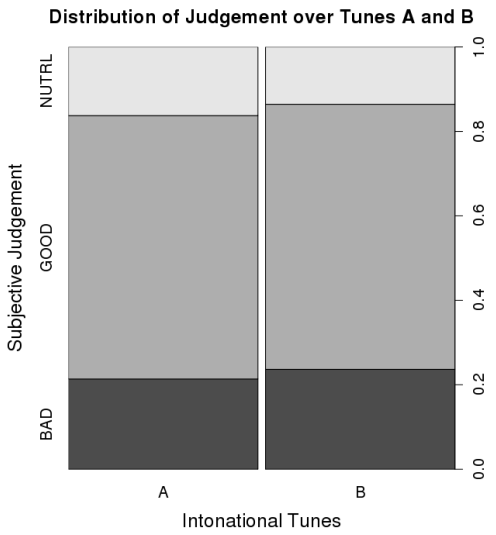


Figure 6.6: Subjective Judgement vs. Intonational Tunes.

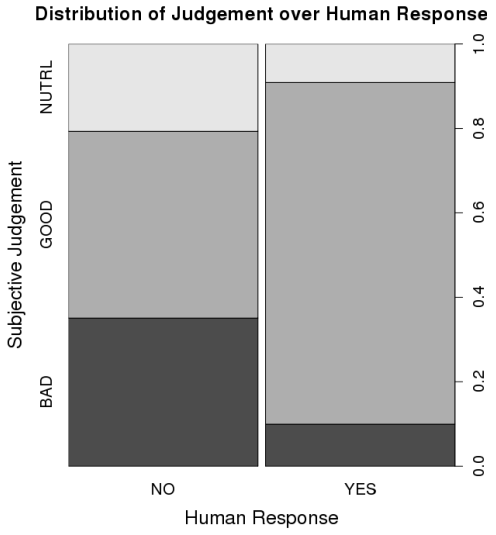


Figure 6.7: Subjective Judgement vs. Human Response.

contradictory to our expectations. This prompts us to analyze the distribution of the score over the third condition, i.e., with regard to whether the robot's hypothesis was correct or incorrect.

For a *correct* hypothesis (YES), tune A is judged as BAD more often in non-congruent condition (23 time) than congruent (13 times). The same applies for tune B which is judged BAD more often (27 times) in the non-congruent condition (NC) than congruent condition (11 times). This is also as expected. However, for an *incorrect* hypothesis (NO), tune A is judged BAD more often (64 times) when in the congruent (C) condition than non-congruent(NC) condition (59 times) (see Figure 6.8). The same also applies for hypothesis tune B, which is judged as BAD more often in congruent condition (80 times) than non-congruent (58 times) (see Figure 6.9). This is contradictory to **Hypothesis 3** and also to **Hypothesis 4**, in which we claim the judgement of tunes to be unaffected by the correctness or wrongness of the robot's hypothesis.

To look into how the robot's hypothesis plays a role in the subjective judgement we investigate **Hypothesis 4**. The plot in Figure 6.7 provide the distribution of the subjective judgement over the human responses ('YES' and 'NO') respectively. It can be inferred from the plot in Figure 6.7 that robot's clarification utterances with human response as 'YES' were judged GOOD much often (595 times) than those with human response 'NO' (297 times). This clearly indicates that the subjects were judging the *correctness* of the robot's hypothesis, rather than judging the appropriateness of the request in context of the visual scene.

The distribution of judgement over the human response clarifies to an extent why we do not see a significant difference between the subjective judgement for

congruent and non-congruent visual contexts (cf. Figure 6.5). As the were judging the correctness of robot's hypothesis they perhaps paid attention to the object being introduced and therefore the presence of other object in the visual context and nuclear accent placement in the intonation did not factor in their decisions. We therefore do not have a very strong evidence on the claim made in **Hypothesis 2**.

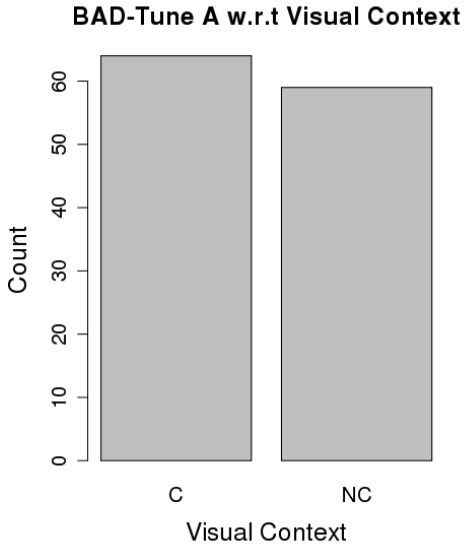


Figure 6.8: BAD judgement for tune A for incorrect hypothesis.

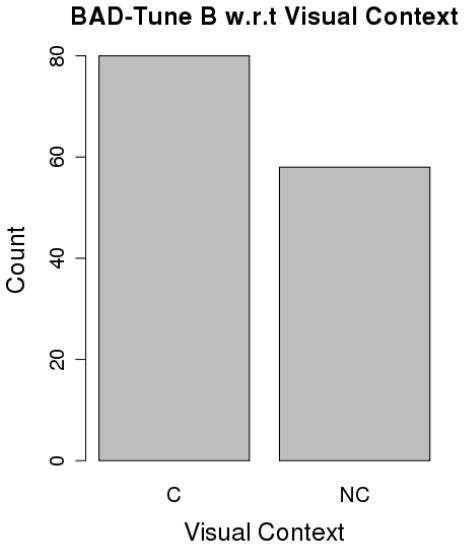


Figure 6.9: BAD judgement of tune B for incorrect hypothesis.

Coming back to the earlier discussion on verification of **Hypothesis 3**, why are the wrong hypothesis judged BAD more often in a congruent case than non-congruent case? On the analysis of the stimuli for these specific conditions (C/NC-A-NO and C/NC-B-NO), we attribute the distinct distribution in Figure 6.8 and 6.8, to the visual context of these stimuli. In a congruent condition the visual scenes offers no ambiguity in the visual context and therefore the subject's visual attention is relatively relaxed, and the decision about the "correctness" of the robot's hypothesis are easier and harsher i.e. judged more often BAD. For non-congruent condition the visual scene offers some ambiguity for the subjects as well, and therefore presumably draws more of subject's visual attention, and perhaps this interferes with the subjective judgement as BAD, i.e. although they were judged BAD because of the "incorrect" robot hypothesis, the visual context lowered the count of BAD for non-congruent cases.

6.4 Discussion and Further Investigations

The preliminary analysis reveals that the acceptability of a clarification requests is influenced by the visual context. With respect to **Hypothesis 2**, we observe that

utterances in which the nuclear accent placement is licenced by the visual context are perceived more often good than the visual context which doesn't licence the accent placement. We do not know of any other similar studies that investigate the role of visual context in establishing the appropriateness of intonation. However, the results seems to provide some evidence on the role of visual context in a situated dialogue.

The findings with respect to **Hypothesis 3** further supports the claim in **Hypothesis 1** that intonational assignment (be it *marked* and *unmarked*) is governed by the visual context. Both marked and unmarked nuclear accent placements are preferred with the visual context licenses it.

The distinctive pattern for the condition (C-NO-A and C-NO-B) provides even stronger evidence on the role of perception of intonation in incorrect hypotheses. The contextually appropriate usage of intonation in incorrect hypotheses leave no scope of ambiguity for the speaker in perception of the speaker's intentions. It can be claimed that an incorrect query in an unambiguous situation is least accepted.

As a further course of analysis, we are planning an eye tracker experiment for verifying if the subjects paid attention to the already present object when making a judgement. We modify the design of this experiment with a change that instead of the human responding to the robot's query that subject would be required to answer the query. In this manner we will be able to involve the subjects in the interaction with the system. Moreover, since the subjects are required to respond to the robot's queries, the objective nature of the task enables us to measure the influence of visual scene and the intonation on their reaction. The hypothesis for this experiment is that with congruent intonation subject will be looking more at the right object, and that they will react faster. At least for the cases where the hypothesis is correct. It's an interesting question whether there will be any differences between the intonation patterns when the robot's hypothesis is wrong.

7 Conclusions

This chapter concludes our thesis. We start by presenting a summary of what we have achieved in this work. We discuss the findings of the ongoing experimental verification of our approach, and outline plans for further investigations. We then provide suggestions for further research on this work.

In this thesis we have presented an approach for realizing *contextually appropriate intonation* in a robot's clarification requests. At the onset of this work, we emphasized with illustrations, that in a situated human-robot dialogue the contextual appropriateness of a robot's clarification requests has to also account for their contextually appropriate intonation. The work presented here follows from our argument that such an appropriateness in robot utterances can be established by accounting for the interplay of a robot's *intention* and *attention*, relative to the *beliefs* it holds.

We have shown how the notion of *information structure* in an utterance's linguistic meaning can be used for presenting this interplay of an agent's belief state and its intentional and attentional state. We base clarification requests and information questions in a *multi-agent belief model* that gives rise to them. We follow Steedman's [2000a] theory of information structure, and represent the *Theme/Rheme*, *Focus/Background*, *Agreement* and *Commitment* informational status of agent beliefs as the four dimensions of IS partitioning.

The novelty in our approach is that we derive these four informational aspects of a belief based on (*i*) whether or not the agent believes/assumes it to be a *common ground*, (*ii*) whether the agent assumes it to be already *salient* or intends to be made salient, (*iii*) whether or not the agent has partial or complete knowledge about the propositional content, and (*iv*) whether or not the agent claims the commitment to know the truth of the belief proposition.

We have also shown the modeling of Steedman's [2000a] Combinatory Categorical Grammar theory, in the OpenCCG framework, for establishing the interface between the information structure semantics and prosodic surface realizations. We have shown that our approach to IS assignment and realization provides an extended model to cover more types of utterances, although in this work we focused particularly on clarification requests and information questions.

The approach has been implemented in a cognitive architecture of a robot employed in a table-top scenario wherein the robot tries to learn a correct model of a visual scene through spoken dialogue with the human tutor. The robot can ask the human for clarification or information if it realizes that it is missing some information or is uncertain about its beliefs. The approach also enables the robot in production of contextually appropriate intonation in assertive responses and acknowledgements.

We have presented a psycholinguistic experimental investigation for verifying the role of visual context in establishing the appropriateness of nuclear accent placement in a clarification request. The preliminary investigations provide evidence for the claim that *marked* usage of nuclear accent placement is preferred when the visual context is congruent to the intonation usage.

7.1 Suggestions for further research

As indicated earlier in Chapter 6, we intend to conduct a series of experiments for verifying the contribution of visual and dialogue context in establishing the appropriateness of contrastive usage of intonation in clarification utterances. The planned line of work is as follows:

1. With the ongoing experiment presented in this thesis, we intend to achieve a baseline indicating the influence of visual context on the subjective judgement of contrastive usage of intonation in clarification requests. Though the results provide some support for our hypothesis, the absence of a explicit evidence commands further investigation. We are currently setting up another experiment using the visual world Eye Tracker. The hypothesis underlying this experiment is that with congruent intonation subjects will be looking more at the right object, and that they will react faster in responding to robot's queries. This allows us to verify the appropriateness of intonation usage in a particular visual context.

2. The ongoing experiment verifies our approach to placement of the nuclear accent in a clarification request. The approach to IS assignment and intonation realization presented here is, however, capable of generating intonation contours with Theme/Rheme phrases, Focus/Background markings, and presentational aspects of speaker's attitude such as Agreement and Commitment. An experimental evaluation of these intonation contours will help us in ascertaining the approach for IS modeling. As the nuances of contrastive theme accents and rheme accents are more subtle then the marked nuclear accent placement, the proposed approach for verification experiment should be as an evolving-context scenario with active interaction with the system.

Besides the experimental verification of the current approach, the work presented in here can be extended further along the following lines of thoughts:

3. The approach for information structure assignment could be extended to determine the contextually appropriate intonation of referential expression [Kelleher and Kruijff, 2006], [Zender and Kruijff, 2007] or verbal descriptions [Zender *et al.*, 2009]. For example, although two discourse entities *e1* and *e2* can be determined to stand in contrast to one another by appealing only to the discourse model and the salient pool of knowledge, the method of contrastively distinguishing between them by the placement of pitch accents cannot be resolved until the choice of referring expressions has been made.

4. The research pursued in this work mainly focuses on contextually appropriate intonation of a clarification request. However, a robot's communication intention for clarification utterance can be realized with various clarification forms [Purver *et al.*, 2001]. It would be interesting to see which alternate forms of an clarification utterance with contextually appropriate intonations are preferred in a context. How does the visual context affect the form and intonation of a clarification request?

List of Tables

List of Figures

A
References

Ades, A. and Steedman, M. (1982). On the order of words. *Linguistics and philosophy*, **4**(4), 517–558.

Adjukiewicz, K. (1935). Die syntaktische Konnexität. *Studia Philosophica*, **1**, 1–27.

Areces, C. and Blackburn, P. (2001). Bringing them all together. *Journal of Logic and Computation*, **11**(5), 657–669. Special Issue on Hybrid Logics. Areces, C. and Blackburn, P. (eds.).

Areces, C. and ten Cate, B. (2006). Hybrid logics. In P. Blackburn, F. Wolter, and J. van Benthem, editors, *Handbook of Modal Logics*. Elsevier.

Areces, C., Blackburn, P., and Marx, M. (2001). Hybrid logics: characterization, interpolation and complexity. *The Journal of Symbolic Logic*, **66**(3), 977–1010.

Baldridge, J. (2002). *Lexically Specified Derivational Control in Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

Baldridge, J. and Kruijff, G.-J. M. (2002). Coupling CCG and hybrid logic dependency semantics. In *ACL'02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Philadelphia, PA. Association for Computational Linguistics.

Bar-Hillel, Y. (1953). A quasi-arithmetical notation for syntactic description. *Language*, **29**, 47–58. Reprinted in Y. Bar-Hillel. (1964). *Language and Information: Selected Essays on their Theory and Application*, Addison-Wesley 1964, 61–74.

Bateman, J. A. (1997). Enabling technology for multilingual natural language generation: the kpml development environment. *Journal of Natural Language Engineering*, **3**(1), 15–55.

Blackburn, P. (2000). Representation, reasoning, and relational structures: a hybrid logic manifesto. *Logic Journal of the IGPL*, **8**(3), 339–625.

Brenner, M. and Kruijff-Korbayová, I. (2008). A continual multiagent planning approach to situated dialogue. In *Proceedings of the LONDIAL (The 12th SEMDIAL Workshop on Semantics and Pragmatics of Dialogue)*. LONDIAL.

Chafe, L. W. (1974). Langugae and consciousness. *Language*, **50**, 111–133.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, **13**, 259–294.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (1999). Minimal Recursion Semantics. An introduction.

Edlund, J., Skanze, G., and Carlson, R. (2004). Higgins - a spoken dialogue system for investigating error handling techniques. In *Proceedings of ICSLP. Jeju, Korea*, pages 229–231.

Edlund, J., House, D., and Skantze, G. (2005). The effects of prosodic features on the interpretation of clarification ellipses. In *Proceedings of Interspeech. Lisbon, Portugal*, pages 2389–2392.

Engdahl, E. (2006). Infromation packaging in questions. *Empirical Issues in Syntax and Semantics*, **6**(1), 93–111.

Firbas, J. (1975). On the thematic and the non-thematic section of the sentence. *Ringbom*, pages 317–334.

Frege, G. (1892). ber begriff und gegenstand. *Vierteljahresschrift fr wissenschaftliche Philosophie*, **16**, 192–205.

Garbe, E., Rosner, B. S., Garciá-Albea, J., and Zhou, X. (2003). Perception of english intonation by english, spanish, and chinese listeners. *Language and Speech*, **46**(4), 375–401.

Halliday, M. A. K. (1967). Notes on transitivity and theme in english, part ii. *Journal of Linguistics*, **3**, 199–244.

Hawes, N. and Wyatt, J. (2010). Engineering intelligent information-processing systems with CAST. *Advanced Engineering Infomatics*, **24**(1), 27–39. To appear.

Hirschberg, J. (1993). Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, (63), 305–340.

Jacobsson, H., Hawes, N., Kruijff, G.-J. M., and Wyatt, J. (2008). Crossmodal content binding in information-processing architectures. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands.

Kelleher, J. and Kruijff, G. (2006). Incremental generation of spatial referring expressions in situated dialogue. In *Proc. Coling-ACL-2006*, Sydney, Australia.

Kruijff, G. (2005). Context-sensitive utterance planning for ccg. In *Proceedings of the European Workshop on Natural Language Generation*, Aberdeen, Scotland.

Kruijff, G. and Janíček, M. (2009). Abduction for clarification in situated dialogue. Cogx year 1 deliverable.

Kruijff, G.-J. M. (2001). *A Categorial-Modal Logical Architecture of Informativity: Dependency Grammar Logic & Information Structure*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

Kruijff, G.-J. M. (2003). Binding across boundaries. In G.-J. M. Kruijff and R. T. Oehrle, editors, *Resource Sensitivity, Binding, and Anaphora*, pages 123–158. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Kruijff, G.-J. M. and Baldridge, J. (2004). Generalizing dimensionality in combinatory categorial grammar. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 191, Morristown, NJ, USA. Association for Computational Linguistics.

Kruijff-Korbayová, I. and Steedman, M. (2003). Discourse and information structure. *Journal of Logic, Language and Information*, **12**(3), 249–259.

Kruijff-Korbayová, I., Ericsson, S., Rodríguez, K. J., and Karagjosova, E. (2003). Producing contextually appropriate intonation is an information-state based dialogue system. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 227–234. ACL.

Lambrecht, K. (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge.

Lison, P. (2008). A salience-driven approach to speech recognition for human-robot interaction. In *Proceedings of the 13th ESSLLI student session (ESSLLI 2008)*, Hamburg, Germany.

Mathiessen, C. M. I. M. (1983). Systemic grammar in computation: the Nigel case.

Monaghan, A. (1994). Intonation accent placement in a concept-to-dialogue system. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 171–174, New Paltz, NY.

Montague, R. (1974). *Formal Philosophy*. Yale University Press.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, Massachusetts Institute of Technology.

Pierrehumbert, J. and Beckman, M. (1986). Intonational structure in japanese and english. Number 3, pages 15–17.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonation in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge MA.

Prevost, S. A. (1996). *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. Phd thesis, University of Pennsylvania, Institute for Research in Cognitive Science Technical Report, Pennsylvania, USA.

Purver, M., Ginzburg, J., and Healey, P. (2001). On the means for clarification in dialogue. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Morristown, NJ, USA. Association for Computational Linguistics.

Purver, M., Ginzburg, J., and Healey, P. (2003). On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*, pages 235–255. Kluwer Academic Publishers.

Rodríguez, K. J. and Schlangen, D. (2004). Form, intonation and function of clarification requests in german task oriented spoken dialogues. In *Proceedings of Catalog '04 (The 8th Workshop on the Semantics and Pragmatics of Dialogue, SemDial04), Barcelona, Spain, July* .

Schröder, M. and Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, **6**, 365–377.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht, the Netherlands.

Skanze, G., House, D., and Edlund, J. (2006). User responses to prosodic variation in fragmentary grounding utterances in dialogue. In *Proceedings of Interspeech ICSLP. Pittsburgh PA, USA*, pages 2002–2005.

Staudte, M. and Crocker, M. W. (2009). Visual attention in spoken human-robot interaction. In *HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 77–84, New York, NY, USA. ACM.

Steedman, M. (1991). Structure and intonation. *Language*, pages 260–296.

Steedman, M. (1996). *Surface Structure and Interpretation*. Essays in Cognitive Psychology Series. The MIT Press, Cambridge MA.

Steedman, M. (2000a). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, **31**, 649–689.

Steedman, M. (2000b). *The Syntactic Process*. The MIT Press, Cambridge MA.

Steedman, M. and Baldridge, J. (2009). Combinatory categorial grammar. In R. Borsley and K. Börjars, editors, *Nontransformational Syntax: A Guide to Current Models*. Blackwell, Oxford.

Stone, M. and Thomason, R. (2002). Context in abductive interpretation. In *Proceedings of EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue*.

Valduví, E. (1990). *The Information Component*. Phd thesis, University of Pennsylvania, Philadelphia, USA.

Vrečko, A., Skočaj, D., Hawes, N., and Leonardis, A. (2009). A computer vision integration model for a multi-modal cognitive system. In *IEEE/RSJ International Conference on Intelligent RObots and Systems*, pages 3140–3147.

White, M. (2004). Reining in CCG chart realization. In *Proceedings of INLG 2004*, pages 182–191. Springer-Verlag Berlin Heidelberg.

White, M. and Baldridge, J. (2003). Adapting chart realization to CCG. In *Proceedings of the Ninth European Workshop on Natural Language Generation*, Budapest, Hungary.

White, M., Moore, J. D., Foster, M. E., and Lemon, O. (2004a). Generating tailored, comparative descriptions in spoken dialogue. In *FLAIRS Conference*.

White, M., Baker, R., and Clark, R. A. J. (2004b). Synthetizing contextually appropriate intonation in limited domains. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*.

Zender, H. and Kruijff, G.-J. M. (2007). Towards generating referring expressions in a mobile robot scenario. In *Language and Robots: Proceedings of the Symposium*, pages 101–106, Aveiro, Portugal.

Zender, H., Kruijff, G.-J., and Kruijff-Korbayová, I. (2009). Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence, Pasadena, CA, United States, AAAI*.

B
Index