February 1996

A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation

Scott A. Prevost
University of Pennsylvania

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-96-01.

A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation

Abstract
In this dissertation I present a model for the determination of intonation contours from context and provide two implemented systems which apply this theory to the problem of generating spoken language with appropriate intonation from high-level semantic representations. The theory and implementations presented here are based on an *information structure* framework that mediates between intonation and discourse, and encodes the proper level of semantic information to account for both contextually-bound accentuation patterns and intonational phrasing. The structural similarities among these linguistic levels of representation are the basis for selecting Combinatory Categorial Grammar (CCG,Steedman1985, 1990a) as the model for spoken language production. This model licenses congruent syntactic, prosodic and information structural constituents and consequently represents a simplifcation over models of prosody developed in syntactically more traditional frameworks.

The *previous mention* heuristic, which has been widely used as a model for determining intonation contours, is shown to be inadequate for handling a broad range of examples involving semantic contrasts, which require pitch accents to be allocated based on their ability to discriminate among available entities in the discourse model. To address this problem, I introduce a model that determines accentual patterns based on sets of alternative entities in the knowledge base. The algorithms for building the information structural representations that encode the semantics of intonation supply the foundation for two computational implementations. These implementations demonstrate how the theoretical model applies to the problem of producing contextually-appropriate spoken output in a natural language generation framework and provide a platform for incrementally testing and refning the underlying theory.

Comments
University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-96-01.

Institute for Research in Cognitive Science

A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation

(Ph.D. Dissertation)

Scott Allan Prevost

University of Pennsylvania
3401 Walnut Street, Suite 400C
Philadelphia, PA 19104-6228

February 1996

Site of the NSF Science and Technology Center for Research in Cognitive Science

IRCS Report 96-01

A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation

Scott Allan Prevost

A Dissertation

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy.

1995

Mark Steedman

Supervisor of Dissertation

Peter Buneman

Graduate Group Chairperson

Acknowledgments

I am deeply indebted to my advisor, Mark Steedman, for motivating this dissertation, and for his patient guidance and thoughtful advice during my years as a graduate student. By his example, Mark taught me what it means to conduct scientific research and provided me with the skills to think like a computer scientist, a linguist and a cognitive scientist all at the same time.

This research has also received tremendous benefit from my interactions with other students and faculty members at Penn and throughout the computational linguistics community. I am particularly grateful to Mark Steedman's research group, which patiently listened to many of the ideas presented here, from inchoate notions to their present form, and to Ellen Prince's discourse seminars, which opened my eyes to the importance of modeling context. Working with Justine Cassell and Catherine Pelachaud broadened my research interests beyond the spoken word, providing insight into other facets of human communication. Matthew Stone, Libby Levison and Beryl Hoffman provided numerous invaluable suggestions and constantly challenged me to think in new ways. Bonnie Webber and Mitch Marcus also provided valuable advice and introduced my work to researchers working in other domains. My piano instructor, David Sokoloff, whose energy defies his eighty-five years, taught me how to dissect problems and listen critically to both music and speech. I am grateful also for the advice and support of my committee members, Ellen Prince, Aravind Joshi, Mark Liberman and Kathy McKeown, whose constructive criticisms guided this research and greatly improved the resulting document.

Without the support of many friends, this dissertation would not have been possible.

I am deeply indebted to Kris Rabberman, who provided the kind of friendship and encouragement that few people are privileged to enjoy, and endured more discussion of my research than any medieval historian rightly deserves. Chris Meek, Dana Izenson and Kim Nagrant, friends from my undergraduate years at Carnegie Mellon, were instrumental in my decision to pursue a Ph.D. My colleagues at AT&T Bell Laboratories provided the encouragement I needed to make the leap into academia. Once in graduate school, the friendship of Kris Rabberman, Tom Misnik, Libby Levison, Matthew Stone and Doug DeCarlo helped me to keep my perspective and stay connected with the world outside of Penn. I am especially thankful for Kris and Libby looking after me in the final days of dissertation writing when I was too busy to sleep and neglected to eat at regular intervals.

Finally, I am grateful for the love and encouragement of my family, particularly my brothers, Drew and David, and my sister-in-law B.J., who have always supported my wildest aspirations and championed my causes. To my partner Robert Kubitz, I owe perhaps the deepest debt of gratitude. His unwavering support, patience and almost infinite tolerance for the demands of doctoral research were instrumental in my surviving graduate school and embarking on such a project. Lastly, I am grateful for the strong influence of my parents, Louis and Isabel Prevost, in whose memory this dissertation is dedicated. Without the knowledge of their love and their confidence in my abilities, this work would not have been possible.

Abstract

In this dissertation I present a model for the determination of intonation contours from context and provide two implemented systems which apply this theory to the problem of generating spoken language with appropriate intonation from high-level semantic representations. The theory and implementations presented here are based on an *information structure* framework that mediates between intonation and discourse, and encodes the proper level of semantic information to account for both contextually-bound accentuation patterns and intonational phrasing. The structural similarities among these linguistic levels of representation are the basis for selecting Combinatory Categorial Grammar (CCG, Steedman 1985,1990a) as the model for spoken language production. This model licenses congruent syntactic, prosodic and information structural constituents and consequently represents a simplification over models of prosody developed in syntactically more traditional frameworks.

The *previous mention* heuristic, which has been widely used as a model for determining intonation contours, is shown to be inadequate for handling a broad range of examples involving semantic contrasts, which require pitch accents to be allocated based on their ability to discriminate among available entities in the discourse model. To address this problem, I introduce a model that determines accentual patterns based on sets of alternative entities in the knowledge base. The algorithms for building the information structural representations that encode the semantics of intonation supply the foundation for two computational implementations. These implementations demonstrate how the theoretical model applies to the problem of producing contextually-appropriate spoken output in a natural language generation framework and provide a platform for incrementally testing and refining the underlying theory.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The melodic characteristics of a spoken language, including the many possible variations in phrasing and pitch, comprise its system of *intonation*. In English, intonational contours carry considerable semantic weight, allowing discourse participants to alter both the meanings of utterances and the manner in which the utterances are interpreted by listeners. In order to produce appropriate intonation in automatically generated synthetic speech, computational approaches must simulate an "understanding" of the significance of the melodies of spoken language. This dissertation addresses this problem by proffering a theoretical framework which links intonational properties to contextual discourse information. A computational implementation which embodies the intonational competence theory demonstrates the appropriateness of the model by automatically generating spoken descriptions of objects with intonation well-suited for the given context.

The principal assumption underlying the theory presented in this dissertation is that intonation is inextricably linked to semantics and discourse context (Bolinger 1972), as demonstrated by numerous examples in the following chapters. The use of stress as a partial indicator of "contrast," which is perhaps the canonical example of the semantic nature of intonation, is examined here with respect to sets of alternative entities and properties put forth by the discourse context. The effects of discourse context on intonation are further complicated by the fact that some languages rely on syntactic phenomena (e.g. Turkish) or morphological phenomena (e.g. Japanese) to convey the types of semantic subtleties produced by the rich intonational system of English. In order to account for

disparities such as these, an intermediate level of information structure, which describes how information in an utterance is packaged with respect to its context, is proposed as a representational bridge between discourse structure and intonation.

An intonational competence model is presented in the framework of Combinatory Categorial Grammar (CCG, Steedman 1991a,1991c,1991b). CCG is a mildly context-sensitive grammatical formalism which licenses congruent syntactic, prosodic and information structural constituents, and consequently represents a simplification over competence models of prosody developed in the syntactically more traditional framework of transformational grammar (Chomsky and Halle 1968, Selkirk 1984). Because of the proposed structural isomorphism among these linguistic entities under CCG, the model for language generation may be considered to involve a single path from high-level semantic representations to spoken utterances, without requiring any additional structures for mapping between surface syntactic forms and intonational domains. The usefulness of this simplified model of speech production is demonstrated by a program that generates paragraph-length monologues with contextually appropriate intonational annotations. The output of the program serves as the input to a speech synthesizer, which in turn produces utterances with the proper intonation.

In this introductory chapter, the motivation for examining for the role of intonation in spoken language generation, and the reasons for adopting the present framework, are given in Section 1.1. Section 1.2 gives a broad overview of research project, essentially providing a road map to guide the reader through the remainder of the dissertation. Finally, Section 1.3 briefly summarizes the principal claims and contributions of the dissertation.

1.1 Motivation for the Present Research

The role of intonation in spoken English is to provide a contextual framework for analyzing the meanings of utterances, for relating the meanings of utterances to one another, and for connecting the belief system of the speaker to that of the listener. Each of these functions of intonation is conveyed by a small set of intonational parameters which indicate phrasing and accentual patterns. For example, consider the following set of question/answer pairs.

(1) a. Q: What kind of music does your older brother prefer?
A: (My older brother prefers) (BAROQUE music).

b. Q: Which baroque art form does your older brother prefer?
A: (My older brother prefers) (baroque MUSIC).

c. Q: Which of your brothers prefers baroque music?
A: (My OLDER brother) (prefers baroque music).

d. Q: Which of your older siblings prefers baroque music?
A: (My older BROTHER) (prefers baroque music).

In these examples, parentheses and capital letters informally indicate intonational phrasing and the most prominent sentential stress (i.e. fundamental frequency (f_0) peak) respectively.[1] While the answers in the four examples contain identical strings of words, they each possess a strikingly different intonation contour. The fact that the contours are not interchangeable without sounding distinctly unnatural provides clear evidence of the effects of context on intonation.

Although several theories have been put forth to explain the accentual patterns in (1), many of these are underspecified for the purposes of accent prediction in synthetic speech, as discussed in Chapter 2. The model that has been employed most effectively in computational approaches makes decisions concerning accent placement on the basis of *previous mention* in the discourse and word class (Terken 1984, Terken and Hirschberg 1994, Hirschberg 1990). This model predicts that open class words are more likely to receive accents, and items that have been previously mentioned in the discourse are likely to be de-accented. While this model is able to account for all of the cases in (1), it fails to account for many instances of *contrastive stress*, such as the answer in (2). In this example, both "baroque" and "music" are explicitly given in the question. Consequently, "givenness" in the sense of "previous mention" cannot possibly account for the decision to accent one and de-accent the other in the answer.

(2) Q: Does your older brother prefer baroque or impressionistic music?
A: (My older brother prefers) (BAROQUE music).

[1] Although secondary stresses are not identified in these examples, the reader should not assume that such features cannot be present.

While the previous examples demonstrate that certain intonational contours sound more natural than others in a given context, the effects of discourse context on intonation are not merely a matter of aesthetics. In the famous example in (3), originally provided by Lakoff (1971), the choice of accentual pattern actually determines the semantic interpretation of the utterance.

(3) a. John called BILL a REPUBLICAN, and then he INSULTED him.
b. John called BILL a REPUBLICAN, and then HE insulted HIM.

The choice of contour in this case not only changes the interpretation of the pronouns, but also makes assumptions about whether or not the speaker believes calling someone a Republican to be an insulting act. Because the notions of intonation and context are so intertwined, as these examples clearly indicate, one can hardly attempt to describe the former without invoking the latter.

One particularly important side-effect of the close relationship between intonation and context is the apparent structural orthogonality of intonational phrasing and traditional notions of syntactic constituency. That is, the demands that context make on intonational groupings of words are often at odds with the demands made by syntax. This problem is illustrated in (4) below, in which the inclusion of the subject noun phrase and the transitive verb in the same prosodic phrase violates the traditional right branching syntactic structure (S $\rightarrow$ NP VP). [2]

(4) Q: What kind of music does your older brother prefer?
A: (My OLDER brother prefers) (BAROQUE music).
L+H* L(H%) H* LL$

As the previous examples demonstrate, the problem of determining intonation contours in context is remarkably complex. Many researchers, including phonologists, semanticists and syntacticians, have addressed this problem, from both purely theoretical and computational standpoints. The primary goal of this dissertation is to address some of the

[2]The intonational annotations utilized in this example are derived from Pierrehumbert's notational system (Pierrehumbert 1980, Beckman and Pierrehumbert 1986, Pierrehumbert and Hirschberg 1990) and are described in detail in Chapter 3. The same chapter includes a discussion of the distinction between intermediate and intonational phrases, a matter which we shall ignore for the present.

deficiencies in the previous approaches by presenting a theory and computational implementation that handles instances of *contrastive stress*, and connects intonation, semantics, discourse structure and syntax in a unified framework.

Given that intonation can be studied in the type of framework briefly outlined above, a key question remains: why endeavor to study intonation from a *generation* and *synthesis* perspective rather than solely by analyzing naturally occurring speech? Although one's deepest insights regarding intonation are precipitated by speech observed in normal conversation, and one can certainly build a theory concerning aspects of intonation merely by experiencing spoken language, one cannot always easily test such theories due to the sparsity of data, interactions with various non-prosodic linguistic phenomena, and the difficulty involved in symbolically representing intonation in transcribed speech.[3] By studying intonation through *generation* and *synthesis*, the theories we develop can be easily and incrementally refined in response to the output of the system. Moreover, the generation task can avoid interactions between intrinsically non-prosodic and prosodic phenomena that are difficult to tease apart in unrestricted natural speech. Finally, in addition to the theoretical justifications for pursuing the present research, there is a particularly compelling practical reason: unnatural intonation is one of the greatest obstacles to winning acceptance of synthesized output in a variety of applications for which the spoken word is a natural method of effective communication.

1.2 Overview of the Dissertation

Many theories concerning the meaning and structure of intonation have been advanced over the past three decades. Chapter 2 is primarily dedicated to examining these models from both theoretical and practical standpoints. In the analysis of theoretical models of intonation, I follow Bolinger(1972), Schmerling (1976), Gussenhoven (1983a) and Selkirk (1984) in arguing that intonation is linked to semantics. An analysis of their theories, however, reveals that none of them is particularly well-suited to handle contrastive stress

[3]Indeed, in a recent workshop on Prosody in Natural Speech at the University of Pennsylvania, I was amazed at the degree to which participants disagreed about how to notate intonation in transcriptions of a wide variety of discourse examples.

in a computational framework. From a practical standpoint, the chapter examines four approaches to assigning intonational contours in synthetic speech (Davis and Hirschberg 1988; Hirschberg 1990; Monaghan 1991; Zacharski *et al.* 1993). These approaches rely on heuristics based on the notions of word class and previous mention. Although such heuristics have proven quite effective for the text-to-speech task, they fail to account for many contrastive accentual patterns.

Chapter 3 lays the groundwork for describing the relationship between the meanings of intonational contours and the discourse structure. The first four sections of the chapter address a range of issues concerning intonational patterns and their meanings, particularly with respect to the issue of referential contrastive stress. Section 3.5 describes the *information structure* level of semantic/discourse representation, which bridges the gap between discourse context and intonation. The chapter concludes with a discussion of the relationship between information structure and local discourse coherence.

In addressing intonation, the present research follows Pierrehumbert (1980) and Beckman and Pierrehumbert (1986) in describing f_0 contours with respect to their constituent parts, including pitch accents, phrasal tones and boundary tones. *Pitch accents* represent relative minima and maxima in the f_0 contour, and generally have the effect of emphasizing the lexical items on which they occur. *Phrasal* and *boundary tones* delimit intermediate and intonational phrases respectively, a matter which warrants some discussion in Section 3.2.3. One crucial difference between the notation devised by Pierrehumbert and the notation adopted here is the granularity of intonational phrase breaks. In Pierrehumbert's system there is no way to specify that some phrase breaks are more pronounced than others (e.g. sentence breaks as opposed to clauses set apart by commas). For the purpose of intonation synthesis, however, such distinctions are crucial. In order to account for varying degrees of pausing and phrasal lengthening, the present approach loosely adopts *break indices* from the ToBI (Tones and Break Indices) annotation system (Silverman *et al.* 1992).

The issue of contrastive stress is addressed in Section 3.4, in which a number of examples are used to demonstrate the weaknesses of the aforementioned previous mention model. An analysis of intonation on pronouns in naturally occurring speech provides

further evidence for this assertion.

Section 3.5 focuses on information structure, which forms the representational link between intonational patterns and the discourse model. That is, information structure packages the semantic content of an utterance with respect to both the context in which it is spoken and the shared knowledge of the interlocutors. The section reviews some of the information structure formalisms that have been offered by other researchers, and introduces the two-tiered structure adopted for the present research illustrated in Example 1.1.[4] The higher tier articulates the theme/rheme structure, borrowing terms from Halliday (1970) and Steedman (1991a). The theme corresponds to the part of the utterance that links it to previous utterances (the "presupposition" in Jackendoff's (1972) terms), while the rheme corresponds to the part of the utterance that forms the core of the speaker's contribution (Jackendoff's "focus"). *Narrow* (or *phonological*) foci within themes and rhemes are represented in the second tier.

Example 1.1

Q: I know the AMERICAN amplifier produces MUDDY treble,

(But WHAT kind of treble) (does the BRITISH amplifier produce)
L+H* L(H%) H* LL$

A: (The BRITISH amplifier produces) (CLEAN treble.)
L+H* L(H%) H* LL$
theme-focus *rheme-focus*
Theme *Rheme*

Steedman (1991a) and others have proposed that the intonational tunes accompanying the utterances in Example 1.1 are associated with the underlying information structure. Specifically, the **L+H* LH%** tune accompanies the *theme* of the sentence, marking what the discourse participants have agreed to talk about. The **H* LL%** tune accompanies the *rheme* of the sentence, marking what is interesting or new about the theme. Although it is certainly unreasonable to assume that these associations hold steadfast for all types

[4]The knowledge base used by the computational implementation described in Chapter 6 contains information about fictitious stereo components. Consequently, many of the examples presented in this dissertation allude to such objects.

of discourse, they seem appropriate within the restricted paradigm of *wh*-questions and simple declarative statements. While the choice of tunes is clearly significant, the location of the pitch accents within the tunes is equally important. Within an intonational tune, the pitch accents mark the words on which they occur as *focused* in the interpretation of the constituent bearing the tune. Elements of the interpretation may be focused for a variety reasons, including emphasizing their newness or making contrastive distinctions among salient discourse entities.

The connection between information structure and discourse organization is addressed in Section 3.6, principally with respect to centering theory (Grosz *et al.* 1986; Joshi and Weinstein 1981), a framework originally formulated to describe local (intra-segmental) discourse coherence.[5] Although the present research appeals to centering theory for generating coherent text, the model given here differs from the centering model by tracking *propositions* concerning discourse entities rather than simply tracking discourse entities themselves.

Having established the connection between intonation and discourse context in Chapter 3, with information structure acting as the intermediary, Chapter 4 describes the algorithms that process discourse information and build the semantic and information structural representations. The algorithms handle the issue of referential contrastive stress by appealing to sets of alternatives inspired by the discourse model (cf. Rooth 1985), and tracking the succession of themes and rhemes throughout the discourse.

In Chapter 5, the structural congruence between intonation and information structure, which is clearly evident in Example 1.1, forms the basis for the selection of CCG as the syntactic framework for the present research. The chapter describes the CCG formalism in detail, argues for its applicability to the task of describing prosodic phenomena, and proposes several prosodic "grammars" for the class of simple monologues produced by the implementation. The adoption of CCG as the model for the production of intonation and speech is crucial since it is only because of CCG's flexible notion of syntactic constituency that the various levels of representation described above can be processed in tandem. CCG therefore simplifies the computational model by avoiding the need for separate syntactic

[5]The term "intra-segmental" is used in the discourse segment sense rather the prosodic sense.

and prosodic modules to compute the syntactic and prosodic surface forms.

The intonational theory presented here is implemented in a system that generates paragraph-length, spoken monologues concerning objects in a simple knowledge base. In Chapter 6, the process of natural language generation, in accordance with much of the recent literature in the field, is divided into three processes: high-level content planning, sentence planning, and surface generation. Two crucial points concerning the role of intonation in the generation process are emphasized. First, since intonational phrasing is dependent on the division of utterances into theme and rheme, and since this division relates consecutive sentences to one another, matters of information structure (and hence intonational phrasing) must be largely resolved during the high-level planning phase. Second, since accentual decisions are made with respect to the particular linguistic realizations of discourse properties and entities (e.g. the choice of referring expressions), these matters cannot be fully resolved until the sentence planning phase.

The high level content planning phase differs from previous algorithms by incorporating information structure and employing a hybrid of several competing methods of organizing information. A variation of the schemata approach (McKeown 1985, 1986) ensures adherence to certain domain- and genre-specific constraints concerning the presentation of information. Within those constraints, the order in which information is conveyed depends on domain-specific knowledge (as in Sibun 1991, 1992), the communicative intention of the speaker and beliefs about the hearer's knowledge. Finally, an approach based on rhetorical structure theory is used to rearrange propositions in order to make certain rhetorical relationships among them, such as contrasts, salient.

The implementation, which is written in Prolog, is defined for a limited set of monologic tasks. Accordingly, its primary purpose is to test the hypothesis that the representations involved in language production, from abstract semantic propositions to surface phonological and syntactic forms, can be treated as structurally isomorphic. As a consequence of this hypothesis, spoken language generation can be viewed as a straight-line process, as exemplified by the implementation architecture shown in Figure 1.1. The secondary purpose of the implementation is to provide a platform for incrementally testing and refining the competence model of intonation embodied by the CCG categories and

rules. Indeed, the present results are a consequence of employing such a methodology.

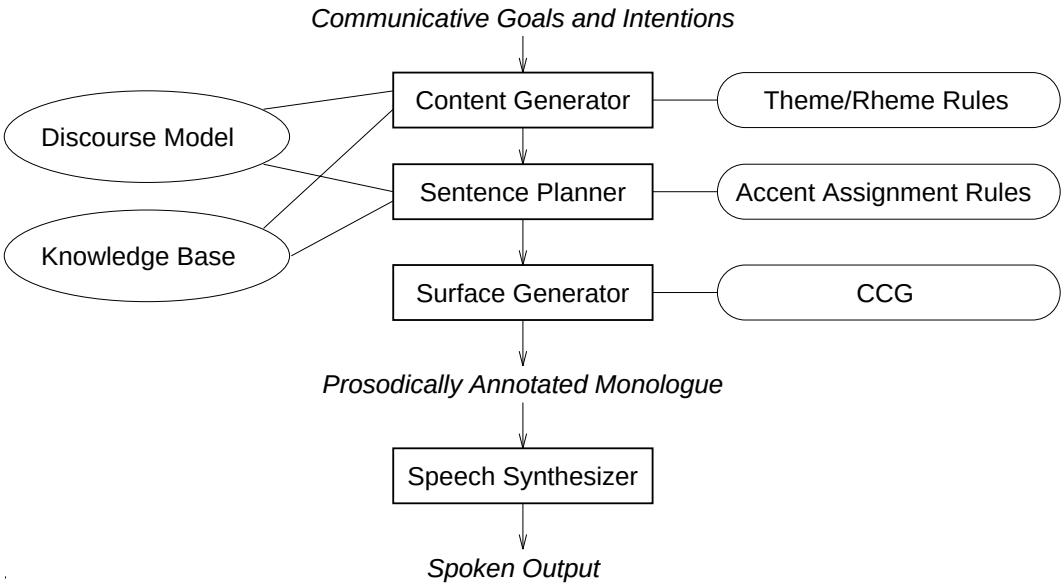


Figure 1.1: An Architecture for Spoken Language Generation

As a natural extension of the present work, the relationship between intonation and other extra-linguistic modes of communication are examined in a computational framework in Chapter 7. This chapter presents an implemented system by Pelachaud and Prevost (1994,1995a,1995b) for animating lip movements and facial expressions in three dimensional graphical models directly from intonational patterns in synthesized speech. The purpose of the exercise is to test hypotheses concerning the synchrony of speech and facial expressions (Condon and Ogston 1971), and to develop an understanding of the relationship of such expressions to discourse context and information structure. The chapter concludes with a brief discussion about how such theories might be formally evaluated.

1.3 Claims and Contributions

In this dissertation, I present a competence model for spoken language production and a computational implementation that generates prosodically appropriate synthesized monologues concerning entities described in a simple knowledge base. The theory conveys proper intonational distinctions of contrast and emphasis with respect to a discourse

model, a domain-independent knowledge base, and the communicative goals and intentions of the modeled speaker. The corresponding program, which produces synthesized speech that varies according to the discourse context, serves as a vehicle for evaluating and refining the intonational competence model.

In the chapters that follow, the discussion of the issues outlined in the previous section provides the framework for advancing the following claims:

i. Any model of intonation that assigns pitch accents based on the distinction between previously mentioned and "new" information does not account for accentual patterns that serve to distinguish among contrasting entities or properties in the discourse.

ii. Certain types of explicitly referential contrastive stress (i.e. accentual patterns that distinguish between two salient discourse entities) can be predicted by a model of *alternative sets* with respect to the discourse context.

iii. *Information structure* effects a mapping between discourse context and intonation and provides a level of semantic detail that accounts for intonational phrasing, accentuation and discourse coherence.

iv. Intonation, information structure and syntax are structurally congruent.

v. CCGs are instrumental for encoding the relationships between intonation, semantics, information structure and syntax.

vi. The problem of generating natural language is facilitated by a bi-level information structure representation. This representation associates the content planning phase with theme/rheme articulations and the sentence planning phase with focal distinctions within themes and rhemes.

vii. Information structure serves as a uniform mechanism for incorporating several competing models of content planning (including schemata, rhetorical structure approaches and domain-driven approaches) into a hybrid framework for the high-level organization of text.

Only the first of these claims is immediately provable. Chapter 3 provides numerous examples that constitute a proof by contradiction. Claim (ii), which is not intended to exhaustively cover all phenomena that have been labeled as contrastive stress, is substantiated by the algorithm in Chapter 4. While the remaining claims are inherently

unprovable, Chapters 3, 4 and 5 explicate the reasons for advancing such hypotheses, and Chapter 6 provides evidence that such claims form the basis for a workable model of spoken language production.

Chapter 2

Literature Review

The task of generating spoken language with context-appropriate intonation encompasses a diverse collection of topics in the field of computational linguistics, including prosodic phonology and its relation to syntax, categorial grammar, focus, information structure, discourse structure, semantic models, and natural language generation. In turn, each of these research areas circumscribe a variety of sub-topics relevant to the present research. To further complicate matters, these topics have been investigated by researchers from a variety of disciplines, including syntax, semantics, phonology, theoretical computer science and artificial intelligence. Each discipline has its own particular concerns which necessarily affect the research methodologies and the presentation of results.

While many of the issues pertaining to this broad collection of topics have been examined in isolation, the research described here attempts to integrate them into a unified framework that can be computationally implemented and tested. The goal of the present chapter is to critically analyze each of these topics in order to provide the reader with the appropriate background information and analytical tools necessary for understanding the theoretical and computational contributions of the subsequent chapters. Because of the wealth of topics covered, it is certainly not possible to present a complete review of all relevant research for each topic within the space of a single chapter. The approach taken here is to delve into detailed analyses for the most important contributions, particularly when such descriptions are not conveniently located in the subsequent chapters for which they are directly relevant. Certain topics which are described in detail in later chapters

are treated more briefly here.

The first section of the chapter, which contains a broad overview of the manner in which the aforementioned areas of research relate to the dissertation, identifies relevant portions of the linguistics, computational linguistics and computer science literature. The remainder of the chapter is divided into separate sections for each relevant research topic, concluding with a section which summarizes the most recent attempts at deriving intonation in text-to-speech applications.

2.1 Relevant Research Areas

Research on prosodic aspects of language, and particularly intonational structure, has only flourished in the last twenty years, lagging well behind research in many other areas of linguistics. The slow development of intonational theories can most likely be attributed to the observation that prosodic structures often appear to be orthogonal to traditional notions of syntactic structure, such as those posited by transformational grammar. Because syntax has been the guiding force in linguistic study and the relationship between prosody and syntax has always been a point of contention, it is not surprising that prosodic aspects of language have received somewhat less attention.

As linguists began to recognize the need for descriptions of intonational structure and meaning, two hypotheses concerning the prosody-syntax connection arose. The first hypothesis, which was clearly driven by the syntax-oriented nature of linguistic research, claimed that syntactic and prosodic structures are indeed the same, except for certain special cases. This was the view put forth by Chomsky and Halle in their 1968 treatise on *The Sound Pattern of English* (henceforth *SPE*), which is rooted in generative grammar (Chomsky and Halle 1968). The SPE approach sought to account for cases of "normal" stress and specifically declined to account for "emphatic" stress, which Chomsky and Halle regarded as the responsibility of the performance theory. Bresnan (1971), Lakoff (1972) and Berman and Szamosi (1972) continued this work on the syntactic modeling of nuclear stress. More recent work by Selkirk (1984) also continued within the framework of generative grammar but eschewed the notion of normal stress.

While much of the research in the SPE mold was widely accepted as correct through the 1970s, in part due to the prevailing syntax-oriented nature of linguistic research, a number of other linguists began exploring the other logical hypothesis concerning the relationship between syntax and prosody. This hypothesis, which claims that syntactic and prosodic structures are not necessarily congruent, accounts for much of the data that proved troublesome to the SPE framework, but suffers from the undesirable property of requiring a more complex cognitive model of language processing. Linguists who accept this hypothesis are forced to accept a model of human language production that handles syntactic and prosodic aspects of speech separately.[1]

In order to address this second hypothesis, some researchers have merely adopted prosodic structures which are distinct from traditional syntactic structures. For example, Nespor and Vogel (1989) define autonomous hierarchical prosodic structures not entirely unlike those introduced by Selkirk (1984). Other researchers have concentrated on semantic aspects of language for defining intonational domains and assigning accents. Most notably, the accounts produced by this latter group include Bolinger's (1972,1989) account of semantic highlighting, Gussenhoven's (1983a,1983b) notion of focus and the Sentence Accent Assignment Rule (SAAR), and Schmerling's (1976) Principles I through IV which relate predicates and their arguments to various stress levels. Much of the literature on English stress patterns is composed of arguments for and against the theories expounded by these three researchers. Still others, such as Terken (1984) and Terken and Hirschberg (1994), have conducted research on the distinctions between *given* and *new* entities in discourse and their relation to intonation. This work retains the semantic bent of Bolinger, Gussenhoven and Schmerling, but also fits neatly into computational frameworks for predicting intonational patterns. Consequently, computational models which are based on the given/new distinction, involving separate syntactic and intonational components, have flourished (cf. Monoghan 1991, Hirschberg 1990).

Because of the syntactic bias in linguistics research, a third hypothesis concerning the relationship between syntax and prosody has until recently been largely overlooked.

[1] To make matters even more confusing, Cooper and Paccia–Cooper (1980) have shown, through a variety of experiments concerning duration, that the role of certain types of syntactic relations in speech processing cannot be ignored altogether.

This hypothesis, which claims that syntactic structure and prosodic structure are indeed congruent, allows for the semantic view of prosodic structure, but also licenses a more flexible notion of syntactic constituency than that licensed by more traditional grammatical formalisms. Steedman (1990a,1990b,1991a,1991c,1991b, 1991d) has been the strongest proponent of this view, and his Combinatory Categorial Grammar (CCG) forms the basis of a unified account of syntax, intonation and information structure.[2] The advantage of such a system, which provides the framework for the present research, is that it allows for a much simpler cognitive model of language processing which does not involve multiple autonomous (and possibly competing) components. The fact that CCG was originally devised to account for purely syntactic phenomena provides evidence that the congruence between syntactic and intonational constituents is not merely convenient, but also linguistically motivated.

Apart from the relationship between intonation and syntax, there are a number of purely intonational issues that are addressed by the present work as well. The primary area of concern is to define what is meant by "intonation," to examine the constituent parts of an intonational contour and to describe how those parts fit together. For this, we rely principally on the work of Janet Pierrehumbert (1980), whose landmark dissertation presented a remarkably complete model of the physical elements of intonation. The notational system devised by Pierrehumbert is widely employed by linguists in the United States and has spawned other systems (e.g. ToBI, Silverman *et al.* 1992) that are closely related.

When phonologists like Pierrehumbert discuss intonational contours, they refer to a wide array of "tunes" (given by f_0, the fundamental frequency contour). Intonational tunes are composed of a variety of different pitch accents and boundaries, all of which possess distinctive physical characteristics, including shape and alignment with lexical stress.[3] The vast majority of the syntactic, semantic and pragmatic accounts of stress neglect such fine distinctions, often resorting to a single notion of stress whose physical

[2]Similar arguments have been made in related categorial frameworks. (Moortgat 1989, Oerhle 1988).

[3]The physical characteristics of pitch contours are clearly manifest in plots of fundamental frequency on the y axis against time on the x axis. Such graphs are generally referred to as *pitch tracks*.

characteristics are left completely unspecified. Even those accounts that incorporate different stress levels (e.g. Schmerling 1976) do not differentiate between different accents in Pierrehumbert's terms. Pierrehumbert and Hirschberg (1990) have filled this obvious gap to some degree by examining the meanings of a variety of intonation contours in discourse. While the analysis is somewhat informal, as necessitated by the complexity of the task, many of their observations are incorporated in the intonation model presented here, as discussed in Chapter 3.

One of the major contributions of this dissertation is the examination of what it means for two discourse entities or propositions to be contrastive and how this can be intonationally encoded. The notion of contrastive stress has pervaded the literature on English intonation but has often been dismissed as a special case outside of what Chomsky and Halle (1968) called *normal stress*. Opinions concerning this phenomenon vary widely among linguists who study intonation, as evidenced by the following:

- Schmerling (1976) objects to classifying any stress as contrastive on the basis that anything can be said to contrast with anything else given the right state of affairs.
- Cruttendon (1986) concurs with Schmerling's assertion that "contrastivity" eludes precise definition, but nonetheless contends that such a notion has linguistic value.
- Couper-Kuhlen (1984) claims that specific intonational contours can be associated with contrast.
- Bolinger (1972,1989) correctly refutes the claim of Couper-Kuhlen.
- Following Dik *et al.* (1980), Gussenhoven (1983a,1983b) discusses contrastiveness with respect to counterassertive and counterpresuppositional sentences.
- Ladd (1980) contends that certain alleged instances of narrow contrastive focus are in fact the result of *broad focus*.

In Section 2.3, after examining these concerns about contrastive stress, we conclude that studying contrastive stress from a purely analytical standpoint has contributed to much of the confusion. Moreover, we argue that the concerns about the elusiveness of the concept

of contrastive stress are not directly applicable to the knowledge- and discourse-based generation model of contrastive stress expounded in this dissertation.

There are two semantic and pragmatic notions of "focus" that are linked to intonation and hence relevant to the present research. First, what has been called *narrow focus*, *phonological focus*, *accent* or *stress* in the literature is crucial for determining minima and maxima in the intonation contour. Second, what has been called *discourse focus*, *theme*, *topic* or (loosely speakly) *backward-looking center* is crucial in maintaining the coherence of discourse from one utterance to another. The former notion of focus (henceforth referred to simply as *focus*) is certainly well represented in the phonology literature by Pierrehumbert (1980), Pierrehumbert and Hirschberg (1990), Selkirk (1984), Bolinger (1972), Schmerling (1976), Gussenhoven (1983a), Ladd (1980), Terken (1984), Beckman (1986), Bird (1991), Cruttendon (1986) and Fuchs (1984), among others. Focus has also been examined in formal semantics by Rooth (1985,1992), Krifka (1992) and Jacobs (1991). While much of the formal semantics work concentrates on focusing particles such as "even" and "only," Rooth's (1985) alternative set semantics, which associates discourse entities with groups of likely alternative entities from the discourse model, forms the the basis for the present model of contrastive stress. Although the focusing particles are not directly addressed in this dissertation, Steedman (1991b) describes them in an intonational framework that is entirely consistent with the one described here.

The second notion of focus alluded to above is related to the information structure of an utterance. Information structure refers to the segmentation of a sentence into separate contiguous parts based on the context of the prior discourse. While a number of different naming conventions have been proposed for the information structural constituents, they all roughly map onto the terminology used by Halliday (1967,1970) and borrowed by Steedman (1991a). Halliday's *theme* roughly marks the cognitive starting point of the sentence, or, in other words, the part of the sentence that links it to the prior discourse (the "presupposition" in Jackendoff's terms). Halliday's *rheme* roughly corresponds to the part of the sentence that the speaker contributes to the discourse (i.e. the new or particularly salient information).[4] Other researchers, such as Hajičová and Sgall (1987,1988), Kuno

[4]The terms "contribution" and "cognitive starting point" are borrowed from Gussenhoven's description of his [+focus] and [-focus] features respectively (Gussenhoven 1983b, p. 18). The reader should note that

(1976), Vallduví (1990), Prince (1986) and Hoffman (1995), make similar information structural distinctions which are described in detail in Chapter 3.

While information structure describes how a single utterance relates to its discourse context, the present research must also be concerned with the overall structure of discourse. That is, we are interested in the linear ordering of sentences as well as the hierarchical ordering of sets of sentences called *discourse segments*. For the former, we appeal to *centering theory* (Grosz, Joshi and Weinstein 1986; Joshi and Weinstein 1981; Walker, Iida and Cote 1990; Brennen, Friedman and Pollard 1987), which was originally conceived as a theory of local discourse coherence. The theory tracks certain discourse entities, called *centers*, through a sequence of utterances and describes the relationship between the centers of consecutive sentences. Hajičová (1987) discusses a similar framework for tracking discourse entities, but additionally considers how such entities may gradually fade from the set of available discourse referents.

Our second concern with the overall structure of discourse, the hierarchical structure above the sentence level, is addressed by the work of Grosz and Sidner (1986), and Polanyi (1988). Due to the limited nature of the types of discourse under investigation in the present research, we are less concerned with the global discourse structure discussed by these authors.[5] For the purposes of discourse generation, however, the research on Rhetorical Structure Theory (RST) by Mann and Thompson (1986) is relevant. RST is a formalism that hierarchically relates phrases and sentences to one another based on a set of rhetorical predicates. Chapter 6 provides a generation account that applies some of the principles of RST to determine discourse organization. Future research is certainly warranted to understand how the discourse segments handled by the present system fit into RST as well as the other hierarchical discourse models.

In addition to the theoretical aspects of the research mentioned above, there are also

---

Halliday made some restrictions on information structure that have been widely (and I believe properly) ignored by others, such as the assertion that themes are sentence-initial.

[5]Unfortunately, much of the literature on discourse structure employs the term "focus" in a manner that is diametrically opposed to the usages by the phonologists and semanticists mentioned above. In the discourse literature, the focus (or center) refers to the information that is at the center of attention, which is generally thematic in nature (in information structural terminology). The information structuralists, on the other hand, would say that it is the rhematic material that is in focus. To avoid confusion, in this dissertation the term "focus" always denotes narrow or phonological focus unless otherwise specified.

a number of computational aspects to be considered, including the problem of generating natural language and the problem of incorporating intonational theories into synthetic speech. The process of generating text has traditionally been divided into two phases: *content* generation, in which high-level goals are satisfied and discourse structure is determined, and *surface* generation, in which high-level propositions are converted into sentences (McKeown 1985). More recently, a number of researchers have proposed an intermediate level of representation, often called *sentence planning*, which maps the high-level abstract semantic representations onto representations that more fully constrain the eventual surface realization. It has been argued by Reiter (1991), Dale and Haddock (1991) and Meteer (1991), among others, that a number of issues in discourse generation, such as lexical choice, are best handled at the sentence planning level. Reiter (1991) provides some compelling arguments for the psycholinguistic plausibility of such an architecture.

The second important computational aspect of the present research is the problem of determining intonational parameters for synthetic speech. The problem of producing speech from prepared texts, which is often referred to as the *text-to-speech* (TTS) task, has been addressed by Monaghan (1991) and Hirschberg (1990) in the framework of given/new distinctions. The more complex task of producing speech from concept, which has been termed the *meaning-to-speech* (MTS) task, has been addressed by Davis and Hirschberg (1988), Zacharski et al. (1993) and previous versions of the present research (Prevost and Steedman 1993a,1993b,1994b).

Having provided a general overview of the significant research topics relevant to the present work, we now turn our attention to detailed discussions of the most important issues stemming from these topics.

2.2 Syntactic and Semantic Accounts of Intonation

There are several possible hypotheses for the relationship between syntactic, semantic and intonational structure, three of which are stated below.

Hypothesis 1 Intonational structure (accentuation) is completely determined by surface syntactic structure (SPE, Chomsky and Halle 1968).

Hypothesis 2 Intonational structure reflects semantic content rather than syntactic structures. Syntactic analyses are inadequate for describing the placement of nuclear (and non-nuclear) stress. (Schmerling 1976; Bolinger 1972, 1989; Gussenhoven 1983a, 1983b)

Hypothesis 3 Both intonational and syntactic structures reflect semantic content and information structure. (Steedman 1991a; Prevost and Steedman 1994b; Moortgat 1989; Engdahl and Vallduví 1994)

The first hypothesis relies on the notion of *normal stress*, which is generally taken to mean the stress that results when focus is unspecified (Ladd 1980) or when the entire sentence can be assumed to be in focus (Gussenhoven 1983a). However, a well-known minimal pair of examples originally given by Newman (1946) seems to defy the notion of normal stress. These examples, shown below in (5) and (6), can both be uttered in the broadest possible context (e.g. as an answer to the question "what's happening?").[6]

(5) I have instructions to LEAVE.
"I've been told to leave."

(6) I have INSTRUCTIONS to leave.
"I plan to leave instructions."

In order to account for such examples, Bresnan (1971) proposed modifications to Chomsky and Halle's nuclear stress rule (NSR) that allowed non-surface syntactic representations to affect phonology. Likewise, Lakoff (1972) and Berman and Szamosi (1972) advanced NSR modifications which maintain a focus on syntactic structure. However, Bolinger (1972) refutes each of these NSR modifications with examples like the following, which seem to defy explanation under any purely syntactic transformations.

(7) I have a point to EMPHASIZE.
I have a POINT to make.

[6]Note that only the word assigned the nuclear stress is marked in uppercase letters in these examples. Other secondary stresses are left unmarked.

(8) The end of the chapter is reserved for various problems to COMPUTERIZE.
The end of the chapter is reserved for various PROBLEMS to solve.

While such examples call into question the rules that have been proposed for determining normal stress, the notion of normal stress itself has received considerable criticism from the Hypotheses 2 camp (including Bolinger 1972, Schmerling 1976 and Gussenhoven 1983a). Their criticism of the notion of normal stress is that it requires the assumption that utterances exist in a contextual vacuum. Schmerling's famous example concerning dying United States presidents, shown in (9), makes this point quite clearly.

(9) Truman DIED.
JOHNSON died.

The first of these pronouncements was uttered at a time when Truman's health was widely known to be quite poor and his imminent death not unexpected. In the second case, Johnson's death came as a surprise because he had not been reported to be gravely ill. The role of context in determining the accentual patterns in these utterances should be clear. Schmerling (1976) also notes that even when subjects are asked to read sentences with as little context as possible, such as those in (10), their placement of stress generally does not support the notion of a "normal" stress assignment. In these latter examples, John and the physics professor are both assumed to be unknown to the reader.

(10) John DIED.
My PHYSICS professor died.

Given that the goal of this dissertation is to generate speech with contextually appropriate intonation, the proposals for handling normal stress are guaranteed, by the very definition of normal stress, to be insufficient for this task. Moreover, the counterexamples posed to the modified NSR accounts by Bolinger and Schmerling indicate that Hypothesis 1 is untenable even as a foundation for the present work. It is for these reasons that we turn to Hypothesis 2.

Hypothesis 2 is based on the belief that accent placement cannot possibly be lexico-syntactic in nature. Of the proponents of this hypothesis, Bolinger leans furthest toward the belief that the process of accentuation is purely semantic and pragmatic in nature. His

account of English stress is perhaps best summarized by the clever journal article title "Accent is predictable (if you're a mind reader)" (Bolinger 1972). Bolinger essentially contends that accented words mark the "information focus" such that less predictable words are more likely to be intonationally highlighted. For example, in (8) the claim is that problems are more likely to be "solved" than "computerized." Bolinger further relates his assertion that words are accented depending on their contextual "interest" to Sperber and Wilson's (1986) notion of "relevance for the hearer" (Bolinger 1989, pp.357–59). In making the comparison, Bolinger appeals to a scale of "interest" which relates accents to their propensity to produce certain contextual effects for the hearer.[7]

While Schmerling (1976) shares Bolinger's view that Hypothesis 1 cannot be correct, she takes exception to the assertion that relative semantic weight accounts for accentuation patterns. In particular, she cites examples of unstressed verbs that are clearly unpredictable and stressed verbs that seem reasonably predictable from context, as shown in (11) and (12) respectively.

(11) Hey—your COAT'S on fire!

(12) Come on in—the door's OPEN.

Schmerling's account of sentence stress is based on the following four principles (Schmerling 1976):

I. Certain items in an utterance are treated by the speaker as relatively "insignificant" and fail to be assigned stress.

II. The verb receives lower stress than the subject and the direct object, if there is one; in other words, predicates receive lower stress than their arguments, irrespective of their linear position in surface structure.

III. Given a sequence of stresses which are equal and greater than other stresses within the intonational unit, the last such stress will be more prominent than the others.

[7]For the purposes of the present work on contrastive stress, this formulation by Bolinger is important because it demands the information in the foreground to produce contextual effects in the hearer's model, even if that information is not "new" to the hearer in any sense.

IV. In a topic-comment utterance, stress both the topic and the comment.

Schmerling claims that the first three of these principles are applicable to "news" sentences (or sentences with broad focus in the terminology of Ladd 1980). Principle II, however, has been strongly challenged by Gussenhoven (1983a) and Selkirk (1984), who both object to this disregard for focus structure. Although Principle IV does address focus structure, it suffers from the Hallidaian assumption that topics (or themes) are utterance-initial. Schmerling accounts for the fact that comments generally receive more emphatic stress than topics by applying Principle III immediately after applying Principle IV in non-news sentences. However, for utterances that are comment-initial, this strategy clearly fails.

An alternative approach taken by Gussenhoven (1983a,1983b) is entirely based on "focus domains." He notes that a single accent in a focus domain can often identify the focus structure of multiple constituents, where the focus structure for an utterance is defined by an assignment of the feature [+focus] or [-focus] to each constituent. A set of domain and accent rules (Sentence Accent Assignment Rules) produces a segmentation of the focused utterance into focus domains and assigns accents within the domains. The Sentence Accent Assignment Rule is formulated as shown below, where the [+focus] feature is denoted by underlining, square brackets delimit focus domains, and asterisks signify accents. A,P and C stand for argument, predicate and condition respectively. X and Y are variables over these categories (Gussenhoven 1983a).

Domain Assignment	$\underline{P}(X)\underline{A}$	$\rightarrow$	[P(X)A]
	$\underline{A}(X)\underline{P}$	$\rightarrow$	[A(X)P]
	$\underline{Y}$	$\rightarrow$	[Y]
Accent Assignment	[ ]	$\rightarrow$	[*]. In AP/PA, accent A.

To support his analysis, Gussenhoven provides a number of examples, several of which are reproduced in (13).

(13) AP → [*AP] Our d*og's disappeared.
ACP → [*A][*C][*P] Our d*og's myst*eriously disapp*eared.
ACP → [*ACP] (Speaking of mysteries) Our d*og's mysteriously disappeared.

While this account is appealing because of its use of focus structure to mediate between the strong semantic account of Bolinger and the rule-based account of Schmerling, it fails to address the relevancy issues posed by Bolinger. This is particularly striking when one examines sentences that contrast items pointedly, such as (14), in which the first line gives the focus structure and the second line the result of applying the SAAR.

(14) a. (Speaking of mysteries) Gilbert loves them, but George hates them.
b. [G*ilbert loves them] but [Ge*orge hates them]

While Gussenhoven has accounted for the focus structure, he fails to consider the relative contextual impact of the focused items. Since the semantic predicates $\lambda x.\lambda y.[loves\ y\ x]$ and $\lambda x.\lambda y.[hates\ y\ x]$ stand in direct contrast to one another, the corresponding lexical items must be stressed.

The approaches taken by Terken (1984) and Terken and Hirschberg (1994) are similar to Gussenhoven's focus account in that they deploy sentence accents based on distinctions of givenness. These techniques, which have been implemented in a number of computational frameworks, also fail to account for apparent instances of contrastive stress. Section 2.5 below reviews some of the speech production systems that rely on the given/new distinction, and Chapter 3 includes a detailed account of why such techniques are unable to account for certain contrastive stress phenomena.

A final point concerning the models that fall under Hypothesis 2 is worth noting. While each of these models attempts to explain the placement of accents, most of them fail to even address the issue of intonational boundaries. As we shall see in Chapter 3, the types of boundaries and their locations do in fact contribute to the overall semantic interpretation of an utterance (cf. also Pierrehumbert and Hirschberg 1990). While it may be possible to interpret Gussenhoven's focus domains as delimiters of intermediate or intonational phrases, the SAAR often predicts focus domains that seem at odds with

the natural intonational phrasing. For example, in (15), the distribution of "quietly" and "buried" into different domains is not consistent with the likely intonational phrasing that groups these words together.

(15) [Truman was QUIETLY] [BURIED] [in INDEPENDENCE] [in 1972]. (Gussenhoven 1983b, p. 28)

Part of the reason for the difficulty in determining prosodic boundaries is the apparent orthogonality of the syntactic and intonational models under Hypothesis 2. It is for this reason that some researchers have adopted Hypothesis 3, in which the syntactic structure of an utterance reflects its intonational structure and information structure.

The idea of integrating information structure into syntactic frameworks has been expounded by Steedman (1991a), Oehrle (1988), Moortgat (1989), Culicover and Rochemont (1983), Rochemont (1986), and Engdahl and Vallduví (1994) among others. Similarly, the idea of integrating information structure into models of intonational phrasing has been put forth by a number of phonologists, including Selkirk (1984), Gussenhoven (1983a), Ladd (1983) and Bird (1991). In order to capture these congruence relations among syntax, information structure and intonational phrasing, Steedman proposes employing a mildly context-sensitive grammatical formalism called Combinatory Categorial Grammar.

Although the formulation of Steedman's Combinatory Categorial Grammar was originally motivated by purely syntactic phenomena, such as non-constituent coordination in English and crossed-serial dependencies in Dutch, both illustrated in (16), its usefulness for handling intonational phrasing for the theme-rheme structure of *wh*-questions was soon evident (Steedman 1991a,1991c,1991b). Examples 2.1 and 2.2 demonstrate just two of the possible intonational phrasing, phonological focus and information structural articulations licensed by CCG for the sentence "the British amplifier produces clean treble."

(16) a. John admired, but Mary detested, the newest member of the team.

b. ...Jan Piet Marie zag helpen zwemman. (from Bresnan et al. 1982)

...Jan$_i$ Piet$_j$ Marie$_k$ saw$_i$ help$_j$ swim$_k$.

"...Jan saw Piet help Marie swim."

Example 2.1

Q: I know the AMERICAN amplifier produces MUDDY treble,

(But WHAT kind of treble) (does the BRITISH amplifier produce?)
L+H* L(H%) H* LL$

A: (The BRITISH amplifier produces) (CLEAN treble.)
L+H* L(H%) H* LL$
theme-focus *rheme-focus*
Theme *Rheme*

Example 2.2

Q: I know the AMERICAN amplifier produces MUDDY treble,

(But WHICH amplifier) (produces CLEAN treble?)
L+H* L(H%) H* LL$

A: (The BRITISH amplifier) (produces CLEAN treble.)
H* L(L%) L+H* LH$
rheme-focus *theme-focus*
Rheme *Theme*

These examples illustrate how CCG, unlike the transformational grammar (Hypothesis 1) formalism discussed above, licenses derivations that reflect various intonational contexts of an utterance. For this reason, the notion of normal stress is completely absent in CCG. Consequently, previous models of intonation generation using CCG make no claims about intonational contours in null contexts (Prevost and Steedman 1993a, 1994b). In fact, unlike the generative grammar models, the CCG formalism itself does not make any predictions about intonational phrasing and pitch accent placement. Rather, CCG provides a framework for realizing intonational decisions made at the semantic, information structural and discourse levels. It is this integration of linguistic systems that is of concern in the present research.

A discussion of the computationally theoretical aspects of CCG, categorial grammars in general, and other mildly context-sensitive formalisms is provided in Chapter 5 below,

which is devoted to the construction of intonational combinatory grammars.

2.3 Contrastive Stress

Contrastive stress is a phenomenon that has received little formal study but a great number of informal mentions in the linguistics literature. The reason for this no doubt lies in the contention that contrastiveness seems to elude formal definition (Schmerling 1976, Bolinger 1961). Nonetheless, it is quite easy to construct examples in which the placement of stress seems to signify to the hearer that two entities or propositions are to be considered alternatives to one another, as demonstrated in (17).[8] In discussing "contrastive" examples like these, there are two issues that must be examined. First, we need to determine whether or not contrastive stress can be differentiated from other types of stress phonologically. Second, we need to investigate the correlation between semantics and contrastive stress.

(17) a. Don't turn LEFT. Turn RIGHT.

b. George HATES BEETHOVEN, but he LOVES MOZART.

c. She didn't watch "M*A*S*H", she watched "KOJAK." (Selkirk 1984, p. 209)

d. John called MARY a REPUBLICAN, and then SHE insulted HIM. (Lakoff 1971)

The theory that contrastive stress is marked by intonational contours that differ from those marking other types of stress has been put forward by Couper-Kuhlen (1984) and soundly defeated by Bolinger (1989). Indeed, one would be hard pressed to say that the general shape of the pitch contour on "he WALKED" is any different in the context of the questions "what did John do?" and "did John walk or run?" Postal (1972) posited a slightly more conservative theory that assigns an absolutely higher level of stress to contrasted items. While Bolinger accepts that "speakers do tend to make such [contrastive] accents more prominent for affective reasons," he contends that it is "the degree to which contrastive accents overshadow other accents" rather the absolute level of stress that is interpreted by the hearer as contrast (Bolinger 1989, p. 354). This writer

[8] In these examples, nuclear stress is marked by capital letters. Secondary accents are marked by small capitals.

tends to agree with Bolinger's assertions and offers example (17)b as further evidence that an absolute level of stress cannot reasonably model contrastiveness. In this example there are two contrastive pairs that are marked by two different levels of stress.[9]

The issue of the relationship between semantics and contrastive stress is somewhat more confusing than the purely phonological issue just discussed. The main concern with the notion of contrastive stress is that it cannot easily be constrained. Citing the examples in (18), Schmerling (1976) asserts that just as "woman" can be taken to stand in contrast to "man," "board" can be taken to stand in contrast to any other physical item. Consequently, she concludes that the "difference [in contrast] is merely one of degree." Bolinger (1961) makes very similar arguments, concluding that the smaller the set of alternatives, the more one is likely to think of stress as contrastive.

(18) a. My candidate for president is a BLACK WOMAN (and not a WHITE MAN).
b. The hammer is over there on top of that BLACK BOARD.

The conclusion that the size of the set of alternatives affects the interpretation of contrastiveness is certainly reasonable. Based on this conclusion, Schmerling further asserts that the entire notion of contrastive stress is "not a particularly useful one" (Schmerling 1976). On this latter point this writer disagrees, attributing the dismissal of the usefulness of contrastive stress to the purely analytical framework within which Schmerling and Bolinger work. That is, they seem to view the problem as one in which the hearer is always forced to reconstruct the (possibly infinite) intended set of alternatives based on the placement of accents. In Chapter 3, I demonstrate that for the purposes of generating language with context-appropriate intonation, it is often necessary to consider sets of alternative discourse entities and propositions when making decisions concerning accent placement, particularly for referring expressions. Furthermore, I show the usefulness of

[9]Having just eschewed the proposition that contrastive stress can be modeled by an absolute degree of stress, it should be noted that the prosodic grammar presented in Chapter 5 may appear to propose what has just been refuted. This is not a contradiction. That is to say that a listener cannot with certainty tell from the height of single pitch accent whether or not the speaker intended the accented item to contrast with some other item. However, in the task of *generating* speech from high level semantic representations, dictating that items semantically marked as contrastive should be realized with a slightly higher pitch accent does not violate Bolinger's assertion concerning relative pitch levels. Furthermore, the prosodic grammar in Chapter 5 accounts for multiple contrasts within an intonational phrase so that each can be marked by different levels of stress.

restricting the sets of alternatives to those entities and propositions explicitly mentioned in the discourse. Finally, in Chapter 4, I present an algorithm which employs alternative sets to produce accentual patterns in generated text.

2.4 Natural Language Generation

The task of natural language generation (NLG) has often been divided into two stages: content generation, in which high-level goals are satisfied and discourse structure is determined, and surface generation, in which the high-level propositions are converted into sentences. While this approach has been employed effectively (e.g. McKeown 1985, 1986 for generating texts about the structure of databases; Paris *et al.* 1987), a number of theoretical issues concerning the plausibility of such a model for the human language processor remain unresolved. Most notably, while some linguistic phenomena appear to be universal across languages and clearly belong to the content module, and other phenomena, such as syntax, possess language-specific features inherent to the surface generation module, there are still other linguistic phenomena that fall somewhere in between. For example, the choice of lexical items, which is clearly language-specific, may be influenced by high-level discourse issues as well. Given the two-way distinction between content planning and surface generation, it remains unclear where issues such as lexical choice should be handled, particularly if one is concerned with constructing a model of human linguistic ability.

Recently many NLG researchers have posited the need for an intermediate generation stage, often termed *sentence planning*, in which high-level abstract semantic representations are mapped onto representations that more fully constrain the possible sentential realizations (e.g. Rambow and Korelsky 1992; Reiter and Mellish 1992; Meteer 1991; Elhadad and Robin 1992; Elhadad 1993; Elhadad, McKeown and Robin 1996).[10] A consequence of such an approach is that the high-level content generator can operate without concern for the grammatical constraints of any particular language and the lower-level surface generator can operate without concern for global discourse structure. Issues often addressed by the sentence planning level include lexical choice (Reiter 1991; Smadja and

[10] Reiter (1994) provides an informative introduction to the problem of sentence planning.

McKeown 1991; Elhadad and Robin 1992; Elhadad, McKeown and Robin 1996), construction of referring expressions (Dale and Haddock 1991), and descriptive explicitness (e.g. "the green car" vs. "the car is green," Meteer 1991). Although the inclusion of this phase has been based primarily on its computational advantages, such as reducing the need for excessive feedback between high and low level generation stages, Reiter (1994) has also argued for the psycholinguistic plausibility of such an architecture.

For the present research on the role of intonation in spoken language generation, two points must be considered. First, since intonational phrasing is dependent on the division of utterances into theme and rheme, and since this division relates consecutive sentences to one another, matters of information structure (and hence intonational phrasing) must be largely resolved during the high-level planning phase. Second, since accentual decisions are made with respect to the particular linguistic realizations of discourse properties and entities (e.g. the choice of referring expressions), these matters cannot be fully resolved until the sentence planning phase.

Another issue that has been the focus of some debate in the natural language generation community is the methodology for organizing propositions in the content generation phase. The schemata approach espoused by McKeown (1985) involves selecting a proper schema for conveying the desired message and fleshing out the details based on the available information. Because the templates are essentially written as regular expressions with optional and potentially repeatable parts, there is a great deal of flexibility in the resulting text. While this approach has proven quite effective in practice (e.g. Rambow and Korelsky 1992, Paris *et al.* 1987), there are some inherent limitations to its use. For example, a given set of schemata may not apply equally well across various domains. Moreover, there is little room for replanning in cases where text is not communicated effectively the first time and the reader (or hearer) desires clarification (Hovy 1992).

Although the schemata approach remains the most popular, several researchers have proposed using Rhetorical Structure Theory (RST, Mann and Thompson 1986) as the basis for content determination and organization. RST is a recursive framework based on a set of rhetorical predicates that represent relations between various levels of information (e.g. clauses, sentences) in text. Each rhetorical relation contains a nucleus,

which includes the primary material, and a satellite, which contains the supporting material. By formulating RST relations as plans which include subgoals as "growth points," Hovy (1988,1993) recursively builds text structure trees. The output of Hovy's system is produced by traversing the tree depth-first from left to right and imposing syntactic constraints along the way. This incremental realization of hierarchical structures and depth-first traversal is reminiscent of earlier work by McDonald (1986) on "description directed control."

Another novel approach to the high-level generation planning problem is offered by Sibun (1991,1992). In her approach, plans for describing architectural floor plans are generated without producing discourse trees. Propositions are linked to one another not by rhetorical relations or pre-planned templates, but rather by the structure of the domain, including physical and spatial properties represented in the knowledge-base. To the extent that the structure of text relates to the structure of the entities one talks about, discourse coherence is guaranteed by such an approach.

Once the high-level content generation and mid-level sentence planning processing have completed, a surface generation phase must translate the semantic representation into a surface string, as described in work by Gerdeman and Hinrichs (1990), Shieber and Schabes (1991) and Hoffman (1995) among others. The model of tactical generation proposed here and described in Prevost and Steedman (1993a) shares many similarities with the Shieber and Schabes model for Tree Adjoining Grammars (TAGs), particularly in the manner in which the syntactic and semantic representations are tightly coupled.

2.5 Text-to-Speech Research

Many of the early attempts at producing intonation in synthesized speech relied on innovative algorithms for producing "default" or "normal" intonation contours without according a prominent role to syntactic, semantic or discourse level constraints (O'Shaughnessy 1977, Allen *et al.* 1987, Sagisaka 1990). While the results from such attempts have been quite impressive, such algorithms are unable to appropriately alter the intonational properties of the synthesized speech to accommodate various discourse contexts. Other

researchers have attempted on many occasions to build text-to-speech (henceforth TTS) systems that produce contextually "natural" intonation with varying degrees of success. Early work on TTS systems, such as Young and Fallside's *Speech Synthesis from Concept* (1979), expounded the significance of information other than the actual text for producing appropriate intonation. Although the Young and Fallside approach did not employ some of the sophisticated techniques for modeling context described below, its use of syntactic structure alone constituted a major first step. The remainder of this section describes some of the advances in this field of research, specifically focusing on recent work that considers semantics and discourse structure in addition to syntactic and orthographic features for determining accent and boundary locations.

The Instituut voor Perceptie Onderzoek (IPO) in the Netherlands has conducted ongoing research on intonation synthesis for the past two decades. Much of the early work from IPO by 't Hart *et al.* was rooted in perceptual experiments whereby listeners were asked to determine similarities between f_0 contours generated by an intonational grammar ('t Hart and Cohen 1973, 't Hart and Collier 1975). The grammars were created by abstracting pitch movements from a corpus of f_0 contours. More recently, Dutch researchers (e.g. Terken 1984) have concentrated on describing the function of intonation with respect to focus and the given/new distinction. This latter line of research is concerned with determining relative levels of givenness for discourse entities and applying pitch accents accordingly. Work in this area has been recently continued by Davis and Hirschberg 1988, Hirschberg 1990, Monaghan 1991 and Zacharski *et al.* 1993.

The relationship between the givenness hierarchy (Prince 1981) and the accentability of lexical items clearly establishes the importance of context in determining intonation. Since contextual cues can be drawn from syntax, semantics and pragmatics, it is clear that systems that exploit the contextual information in these linguistic areas are bound to produce more natural intonation than systems that rely solely on orthographic cues. Of course the scope of language to be generated is constrained by the limited computational access to syntactic, semantic and discourse structures. Consequently, most TTS systems fall into two categories: those that rely on scant contextual information but handle unrestricted or loosely-restricted text, and those that employ syntactic, semantic and

pragmatic relations to produce utterances in a tightly constrained domain. The remainder of this section describes two systems for pitch accent assignment in unrestricted text, and several attempts at limiting the domains of those systems to gain contextual knowledge.

Julia Hirschberg's NewSpeak system, an interface to the AT&T Bell Laboratories TTS system, is designed to assign pitch accents, pausal duration, speaking rate, and phrasing in unrestricted text based on syntactic information and discourse structure information inferred from orthographic cues (see Hirschberg 1990). The inclusion of a discourse model, based on the Grosz and Sidner (1986) model, allows issues of accentability to be determined with respect to givenness within a discourse segment, thereby incorporating contextual information. The paucity of semantic and pragmatic information, however, restricts the notion of givenness to lexical givenness (based on word roots). Consequently, the system is likely to improperly accent items whose givenness is established by synonymy or implication. More importantly, such a system is completely incapable of assigning accent on the basis of semantic contrastiveness. Unfortunately, these types of errors seem to contribute immensely to the perception of unnaturalness in synthesized speech, as noted by the author.

Regardless of the drawbacks inherent in the NewSpeak system, the amount of intonational information that can be correctly specified given the crude syntactic and orthographic cues makes the system worth exploring. The discourse structure of NewSpeak is composed of a stack of local focus spaces which are pushed and popped based on the discourse segmentation inferred from punctuation, paragraphing and cue phrases. Each focus space contains the roots of open-class words (i.e. content words such as nouns and verbs). A global focus space stores the roots of open-class items that occur in the topic sentence (assumed to be the initial sentence). The algorithm accents all open-class items except for those occurring in the local or global focus spaces (those that are already "given"), and those that are normally de-accented as part of a nominal compound.

The flaws in the NewSpeak system are clearly due to the fact that the algorithm assigns pitch accents based on word class and lexical givenness. Several advances over the NewSpeak model of reading unrestricted text are offered in the more recent work

by Monaghan (1991) described below. Ironically, many of the problems inherent in handling unrestricted text were predicted by Davis and Hirschberg (1988) two years before the NewSpeak system was produced. In describing the Talker module of the Direction Assistance program, Davis and Hirschberg allude to the deficiencies of unrestricted TTS systems by making the following observation:

> While text-to-speech synthesis must rely primarily upon structural information to determine appropriate intonational features, speech synthesized from an abstract representation of the message to be conveyed may employ much richer sources. (Davis and Hirschberg 1988)

Direction Assistance is an application that determines a route and provides spoken directions for traveling between two points on a Boston road map (Davis and Hirschberg 1988). Because of the limitations of the domain, the ability to assign some type of semantic representation to the utterances becomes readily apparent. Consequently, issues of givenness can be resolved based on semantic factors rather than purely lexical factors. Furthermore, the fact that the system generates the text allows for complete control over the discourse structure, thereby avoiding the problem of inferring discourse structure from textual cues. Unfortunately, the system makes only limited use of the available semantic and discourse level information, completely ignoring the notion of contrastive focus.

Direction Assistance works in the following manner. After a route, composed of a sequence of *acts*, is mapped out, a set of descriptive schemas (templates with empty slots) are filled out with values from the abstract *route* representation. The discourse segmentation structure is determined from *discourse segment purpose* relationships specified within the schemas. The discourse segments have associated focus spaces containing sets of potentially accentable discourse entities along with possible semantic annotations. As in the NewSpeak program, pitch accents are assigned based on givenness with respect to the focus spaces. That is, an open-class item occurring in the text is accompanied by a pitch accent unless the item is "given" either lexically or semantically in the focus space.

In addition to determining pitch accent locations, the Direction Assistance program uses discourse segmentation and embedding information to determine pitch range, final

lowering, and pausal duration between discourse segments. While these aspects of intonation are clearly important, the naturalness of the spoken directions would probably benefit more from an improved algorithm for pitch accent assignment that makes use of semantic information to convey contrastive relations. Such a model would certainly be useful in a domain where stark contrasts often have great significance (e.g. FIRST street vs. SECOND street, turn LEFT vs. turn RIGHT, AFTER the bridge vs. BEFORE the bridge, next to the BLUE house vs. next to the WHITE house).

The NewSpeak and Direction Assistance projects share many similarities, but differ in the availability of semantic information. Researchers at the University of Edinburgh have taken a similar approach, developing models for specifying intonation in both unrestricted text (INTERFIX) and highly constrained text (BRIDGE) with rich semantic representations (Monaghan 1991, Zacharski et al. 1993). Consequently, many of the issues addressed by Davis and Hirschberg, relating to the differences between the restricted and unrestricted cases, are also addressed by the Edinburgh researchers.

The INTERFIX program, designed by Monaghan (1991), is similar to NewSpeak in that it attempts to read unrestricted text with reasonable intonation. The goal of this approach, however, is not necessarily to produce the most natural intonation, but to produce *acceptable neutral* intonation. *Acceptable* intonation is defined as intonation which is possible in the given context, but not necessarily the most natural. *Neutral* intonation is defined as intonation that "deliberately leaves the focus structure as ambiguous as possible" and "depends on factors other than syntax" (Monaghan 1991). Monaghan specifically notes that neutral intonation excludes contrastive stress phenomena. The basic idea underlying the term *acceptable neutral intonation* is to produce prosodic domains that sound reasonable in a vacuum, such that a listener can easily construct a context for which the prosody is highly acceptable. Monaghan correctly notes that humans are able to read text with reasonable intonation without necessarily comprehending it. The INTERFIX system claims to model such *naive* intonation.

INTERFIX works by first identifying phonological domains (also called tone-groups, or simply TGs) which specify the locations of boundaries. Three types of TGs are identified: major syntactic constituents (e.g. noun phrases and verb phrases), full clauses,

and complete sentences. The major syntactic constituent boundaries (tg(0)) are obviously gleaned from syntactic analyses. The intermediate domain boundaries (tg(1)), such as parentheticals and lists, and the sentential boundaries (tg(2)) are determined on the basis of punctuation. The types of information employed to determine the phonological domains are similar to the kinds of information utilized by NewSpeak to determine discourse structure. Monaghan's domains, however, lack the hierarchical structure exhibited by NewSpeak's discourse segmentation.

The algorithm employed by INTERFIX for specifying pitch accent locations is reminiscent of the NewSpeak rules, although somewhat more elaborate and less dependent on context. Two levels of stress are assigned on the basis of several rules. First, some content words (e.g. nouns and proper names) are assigned primary stress, while others (e.g. main verbs, adverbs and adjectives) are assigned secondary stress. Function words receive no stress. Clearly such a strategy assigns far too many accents, a fact which is rectified by the *Domain-General Rhythm Rule* which advocates, within a tg(0) domain,

1. deletion of all accents to the right of the rightmost primary accent,
2. reduction of all primaries, except the rightmost, to secondaries, and
3. deletion of every other secondary to the left of the primary. (Monaghan 1991)

Finally, INTERFIX applies a set of *Well-Formedness Conditions* to all phonological domains to ensure that necessary accents are not inadvertently deleted. Whereas NewSpeak deletes potential accents on the basis of lexical givenness, INTERFIX deletes potential accents within prosodic domains on the basis of rhythmic alternation.

Monaghan's intonation assignment algorithms described above form the basis for the intonational component of the BRIDGE (Basic Research on Intonation for Dialogue Generation) project currently under development at the University of Edinburgh (Zacharski *et al.* 1993). Like the Direction Assistance project, the goal of the BRIDGE project is to produce more natural intonation by constraining the application domain, generating text from scratch rather than simply reading pre-written text, and incorporating more elaborate semantic and discourse-level structures. Unlike Direction Assistance, BRIDGE is designed to work with computer-generated dialogues rather than monologues.

The BRIDGE project relies on a program (JAM) which produces dialogues for a map task project, where an agent with a map tries to communicate a map route to another agent with a similar (although not necessarily identical) map. The linguistic issues concerning shared knowledge for the map task are quite similar to the issues in Houghton and Pearson's system described below. (Houghton and Pearson 1988). Because the dialogues produced by JAM are restricted to the map domain, the semantics and discourse structures are readily available and need not be inferred by the module that assigns accents and boundaries. Contextual information is represented as a collection of cards in a *file* system, where a history of previously conveyed information is stored. The file system can be utilized in a manner similar to the focus spaces of NewSpeak and Direction Assistance to resolve issues of givenness.

Due to the variety of available discourse moves (e.g. declarative utterances vs. questions) in the map task, the choice of intonational tune is perhaps as important as the location of accents. By mapping tune choices onto potential discourse moves, the BRIDGE system exhibits many similarities with the system described in the present proposal. We shall see in later sections how the type of question posed to a query system affects the choice of the appropriate tunes for both the question and response.

Boundaries and accents are assigned by the BRIDGE program in a manner similar to that described in Monaghan's 1991 system, with the additional participation of semantic cues and contextual givenness information. Specifically, the four factors which determine accent placement in BRIDGE are linear order preference for rightmost items, preference for content words over function words, preference for semantically "heavy" content words over "empty" content words, and preference for items that are not contextually given.

In work that predates the BRIDGE project, Houghton, Isard and Pearson undertake a similar task of assigning intonation to computer-generated, goal-directed dialogues (Houghton and Pearson 1988, Houghton 1986, Houghton and Isard 1987). The dialogues involve two agents who need to cooperate with each other to manipulate objects in a world consisting of doors, locks and blocks. Boundaries are produced in the dialogue by transforming the syntactic trees into flatter *accent domain* trees by applying recursive rules for deleting and raising certain nodes. Accentability decisions are then determined

on the basis of syntactic word class.

The main tenet underlying the wealth of TTS research described above is that semantic and pragmatic information is vital for producing acceptable and natural intonation contours in synthesized speech. The present proposal adopts this basic principle and attempts to apply it to an even greater extent than any of the systems discussed above. Specifically, the system described in this proposal will make decisions concerning pitch accent placement not merely on the basis of semantic givenness, which has been shown to be inadequate, but also on the the basis of contrastiveness. Neither the Direction Assistance project nor the BRIDGE project have attempted to make such distinctions despite clear evidence that contrastive stress is relevant to their map related application domains. Moreover, the methods employed by these systems for determining prosodic domains (intonational phrases) are often somewhat ad-hoc, based primarily on syntactic and orthographic cues rather than a discourse structure model like the one advocated in our system. Consequently, although our system handles a rather limited type of dialogue, we believe our model is more likely to produce prosodically natural accents and boundaries.

2.6 Summary

This chapter includes an overview of the major issues concerning the formal theories and computational models of spoken language production. From the standpoint of linguistics, the issue of the relationship between prosody and syntax has been a matter of major disagreement. The evidence that intonation is intimately related to semantic and pragmatic aspects of language has caused many linguists to reject associations between intonation and syntax. More recently, however, categorial grammarians have presented syntactic frameworks that intertwine notions of syntactic and semantic constituency, thereby eliminating the need for complex models that map between orthogonal syntactic and intonational structures. Consequently, this latter formulation of the problem, which serves as the foundation for the present research, constitutes an attractive competence theory for spoken language production.

Similarly, much controversy has surrounded the issue of *contrastive stress*. While

there seems to be very little evidence to support the conclusion that there are uniquely contrastive intonational contours, certain patterns of pitch accent placement serve the purpose of identifying discourse entities from among sets of alternatives. The *previous mention* heuristics employed by meaning-to-speech programs are inadequate to account for such accentual patterns. The theory presented here addresses this problem by explicitly modeling alternative set semantics and appealing to notions of information structure and focus.

The task of natural language generation (NLG) has often been divided into two stages: content generation, in which high-level goals are satisfied and discourse structure is determined, and surface generation, in which the high-level propositions are converted into sentences. Recently, many NLG researchers have posited the need for an intermediate sentence planning generation stage, in which high-level abstract semantic representations are mapped onto representations that more fully constrain the possible surface structures. The present work addresses the problem of determining the information structural correlates of intonational features in this three-tiered generation framework.

Chapter 3

Intonation and Discourse

In spoken English, the prosodic characteristics of an utterance determine how the listener perceives its meaning and how the information it conveys is accommodated in the listener's discourse model. The discourse model is the theoretical abstraction that describes the relationships among the various entities, propositions and concepts of the conversation. The connection between discourse and intonation is quite complex, as the examples in the previous chapter clearly illustrate. While those examples focus primarily on the contextual effects of accent placement, decisions concerning intonational phrasing are also affected by the semantic content of prior utterances. The aim of the present chapter is to present an *information structure* formalism that mediates between intonation and discourse, and accounts for both contextually bound accentuation patterns and intonational phrasing.

The chapter begins with a brief introduction to intonational concepts, including a review of the compositional approach to intonational structure and meaning (Pierrehumbert 1980; Beckman and Pierrehumbert 1986; Pierrehumbert and Hirschberg 1990) and an analysis of the use of stress as a partial indicator of contrast. Section 3.5 relates the intonational phrasing and accentual patterns of utterances to their information structure, which defines how information in the utterance is packaged according to context. The manner in which utterances are organized in a discourse and how that organization relates to information structure is examined in the final section of the chapter. Taken together, these relationships among intonation, information structure and discourse form the foundation for the phrasing and accentuation algorithms described in Chapter 4.

3.1 Intonation Contours and Meaning

The relationship between the intonational characteristics of an utterance and the surrounding discourse context is often considered purely from the standpoint of naturalness. That is, for any given context, one accentual pattern can often be said to sound more appropriate than another for conveying a given message. The connection between intonation and meaning, however, is more than a simple matter of aural felicity. In fact, by modifying intonational patterns, one can often completely alter the meaning of an utterance, as illustrated by the accentual patterns in (19)a and (19)c, where the elided verb phrase is ambiguous between the readings in (19)b and (19)d.[1,2]

(19) a. John$_i$ thought BILL$_j$ would SUCCEED, but he$_j$ DIDN'T.

b. *Bill didn't succeed.*

c. John$_i$ thought BILL$_j$ would SUCCEED, but HE$_j$ didn't.

d. *Bill didn't think that Bill would succeed.*

Even in cases where intonational variances do not explicitly alter the semantic content of an utterance, certain variations seem more likely to mislead the hearer than others. For example, given a choice between a bottle of red wine and a bottle of white wine, both utterances in (20) convey the same meaning. By completely de-accenting "white," however, the second response seems to imply that the speaker believes some other alternative to wine to be available. In such a case, one can reasonably argue that the hearer is more likely to add such a belief to her knowledge.

(20) a. I'll have the WHITE wine.

b. I'll have the white WINE.

At the very least, the accentual pattern can affect the relative importance or salience of discourse entities, as illustrated by the nuclear stress placement in the answers in (21) and (22). The importance of context in determining the intonational patterns in these examples is certainly evidenced by the fact that the answers cannot be interchanged without sounding distinctly unnatural, as demonstrated by (22).

[1] As in previous chapters, these examples informally denote intonational contours by capitalizing stressed words.

[2] I owe this example (or a closely related one) to Bonnie Webber. Similar examples were suggested to me by Hardt (p.c., see also Hardt 1993).

(21) Q: What kind of music does GILBERT prefer?

A: Gilbert prefers JAZZ.

Q: Which musician prefers JAZZ?

A: GILBERT prefers jazz.

(22) Q: What kind of music does GILBERT prefer?

A: #GILBERT prefers jazz.

Q: Which musician prefers JAZZ?

A: #Gilbert prefers JAZZ.

One problem with analyzing these inappropriate intonational contours merely as instances of unnatural sounding intonation, however, is that such an analysis fails to consider that discourse involves more than the stream of words exchanged between the interlocutors. Participants in a conversation continually make propositional inferences about the state of the world and the belief systems of their conversational counterparts based on a number of extralinguistic cues, including intonation, facial expressions and gestural movements.[3] Consequently, one cannot assume that unnatural sounding intonation can easily be accommodated in the listener's discourse model. Rather, the intonation contours in example (22) are likely to cause the questioner to consider the possibility that the respondent had somehow misinterpreted the inquiry.

As discussed in the previous chapter, it is precisely these links between intonational contours and discourse context that have caused Bolinger (1972) to expound the semantic highlighting theory, and Gussenhoven (1983a) and Selkirk (1984) to incorporate focal articulations into their intonational frameworks. Before exploring these issues any further, however, it is necessary to examine the physical characteristics of English intonation and settle on a notation for describing such features, the problem to which we now turn.

3.2 Abstract Intonational Descriptions

The intonational descriptions employed throughout this dissertation are are based on the theory of intonation devised by Pierrehumbert (1980) and later modified by Beckman and

[3]Chapter 7 examines the relationship between intonation and facial expressions in a computational framework.

Pierrehumbert (1986). This theory, which describes fundamental frequency in terms of relative prominence, emerged from earlier theories of metrical and autosegmental phonology (Liberman 1975; Goldsmith 1976). In Pierrehumbert's model, an intonational melody, which is given by the fundamental frequency (f_0) contour, is symbolically described by a collection of discrete intonational components, including *pitch accents*, *phrasal tones* and *boundary tones*. This symbolic representation allows the infinite variability in fundamental frequency shape to be mapped onto a finite set of intonational categories. The system is therefore well suited for establishing the connections between classes of intonational contours and their roles in discourse (Pierrehumbert and Hirschberg 1990).

While Pierrehumbert's notational system has been widely employed by phonologists in the United States for over a decade, a slightly modified version of the system has recently been integrated into the ToBI (Tones and Break Indices) transcription system (Silverman *et al.* 1992; Beckman and Hirschberg 1994; Pitrelli *et al.* 1994). ToBI combines the tonal annotations of intonation devised by Pierrehumbert with a level of description concerning break indices (i.e. intonational junctures). Since the present research is concentrated on a restricted number of intonational contours out the many possibilities, most of the subtle differences between the Pierrehumbert and ToBI conventions will not be addressed here. The notion of break indices, however, will be shown to be of crucial importance for the problem of generating intonation. Consequently, after describing the relevant aspects of the Pierrehumbert and ToBI systems, the section concludes with a brief description of a hybrid notational convention for describing tunes and break indices for the generation problem.

3.2.1 Pierrehumbert Notation

Intonation encompasses a variety of prosodic phenomena, including stress, pitch range and phrasing. The model of intonation for English devised by Pierrehumbert (1980) in her doctoral dissertation identified a core set of intonational components from which any intonational contour can be constructed. These components fall into three classes: *pitch accents*, *phrasal tones* and *boundary tones*.[4] There is an appealing analogy between these

[4]Although Pierrehumbert originally referred to the phrasal tones as "phrase accents," this writer and Pierrehumbert herself now adopt the former term because it cannot be confused with the notion of accent

three classes and the grammatical parts of speech in that they both define finite sets of elements that can be combined in an infinite number of meaningful ways. By identifying the ways in which accents, phrasal tones and boundary tones can combine to form f_0 contours, Pierrehumbert has essentially defined the *syntax* of English intonation. This syntax is described in the present section in preparation for the discussion of intonational semantics.

While words in the English language are associated with a simple *lexical stress* pattern that assigns one syllable the greatest prominence, it is the relative prominence of words in an utterance that determines its intonational contour. This latter type of prominence is specified by *pitch accents*, which align certain lexical items in an utterance with relative minima and maxima in the fundamental frequency contour. The two most basic pitch accents are **H*** and **L***. An **H*** accent on a given word is realized as a tone occurring high in the speaker's pitch range whose peak occurs on the lexically stressed syllable of the word. An **L*** accent, on the other hand, is realized as a tone which occurs low in the speaker's pitch range and is similarly aligned with the stressed syllable. In addition to these simple accents, there are four bitonal pitch accents: **L+H***, **H*+L**, **L*+H** and **H+L***. For a word marked by any of these complex accents, the tone marked with the asterisk is aligned with the lexically stressed syllable. The other tone represents the rising or falling nature of the accent. For instance, an **L+H*** pitch accent on the word *computer* would be intoned as a low pitch rising to a distinct high on the second (stressed) syllable.

In addition to the pitch accents, which have the effect of emphasizing certain words, there are two other intonational components, the phrasal and boundary tones, that demarcate phrases. When one of the two *phrasal tones*, **H** or **L**, occurs after a succession of pitch accents, it delimits an *intermediate* phrase. The phrasal tone has the effect of controlling the pitch between the most recent pitch accent and the end of the phrase. A boundary tone, either **H%** or **L%**, can follow an intermediate phrase or a sequence of intermediate phrases to form a full *intonational phrase*. Simply stated, boundary tones describe the general direction (rising or falling) of the f_0 contour at the end of an intonational phrase, and are often associated with some degree of lengthening (i.e. stretching

as stress. As a result, the term "accent" can be used unambiguously to refer to pitch accents.

the contour) and pausing. In actuality, the realization of a boundary tone is affected by the phrasal tone that it follows. Specifically, after an **H** phrasal tone, a process called *upstepping* raises any boundary that immediately follows it (Pierrehumbert 1980).

Since the types of tones introduced above are all described in terms of relative pitch, a mechanism for realizing a melody from sequences of such tones is necessitated. A succession of **H*** accents is realized by a constant high pitch with noticeable dips between accents. For the complex pitch accents involving two tones, a process called *downstepping* compresses the pitch range after each accent. Thus, a sequence of **H*+L** accents would be realized as a descending staircase with the horizontal parts of the wave falling on the stressed syllables. The effects of downstepping are negated by an intermediate phrase boundary.

3.2.2 Tones and Break Indices

As research in prosody (particularly with respect to computational applications) has progressed, the need for a standardized notational system for prosodic phenomena has become apparent. The ToBI system (Silverman *et al.* 1992), which specifies conventions for annotating speech samples with orthographic, tonal and break index information, was designed specifically with this need in mind. Although ToBI's tonal tier represents a slight modification to the Pierrehumbert annotation system, most of the discrepancies between the two systems do not affect the tunes discussed here. The ToBI convention for downstepped high pitch accents, however, has been adopted in the present approach. Rather than employing the **H*+L** accent as described above, downstepped accents are marked with the "!" diacritic. As a result, the sequence **H*+L H* LL%** in the old notation is written as **H* !H* LL%**.

The primary reason for the consideration of ToBI conventions is the inclusion of the break index information, which describes the nature of junctures between orthographic items. Note that in the "pure" Pierrehumbert system, all boundary tones are created equal with respect to lengthening and pausing. Consequently, there is no way to distinguish sentence-medial boundaries from sentence-final boundaries. For the purposes of generating intonation in synthesized speech, such distinctions are crucial. In general,

Break Index	Description
0	Clitics
1	Phrase-Medial Word Boundaries
2	Weak Intermediate/Intonational Phrase Boundary, or Strong Phrase-Medial Boundary
3	Intermediate Phrase Boundaries
4	Intonational Phrase Boundaries

Table 3.1: ToBI Break Indices

sentence-medial boundaries, such as those marked by commas, are associated with a lesser degree of pausing than sentence-final boundaries. The break index tier of the ToBI annotation system, described in Table 3.1 begins to address this situation by associating three different values with intermediate and intonational phrase breaks.

3.2.3 A Hybrid Intonational Notation

Since the task of generating intonation requires a greater degree of granularity in the description of phrasal pauses than the speech annotation task, it will be necessary to include break indices in intonational descriptions used for generation. While the ToBI system described above offers a mechanism that introduces some granularity, it fails to make the precise distinctions required to account for the differences between intermediate phrase boundaries, utterance-medial intonational phrase boundaries and utterance-final phrase boundaries. In order to account for utterance-medial boundaries that are distinctly marked by a boundary tone, it would seem appropriate to have a break index that falls somewhere between indices 3 and 4 on the ToBI scale.[5] Rather than confusing the issue any further by interjecting new indices into the ToBI classification system, the following notation is used to differentiate full intonational phrase boundaries. The diacritic "%" is used to mark utterance-medial boundaries, as for phrases set apart by commas, and the diacritic "$" is employed for utterance-final boundaries.

It should also be noted that although Pierrehumbert makes a distinction between intermediate and intonational phrases, the inclusion of break indices in the intonational

[5]While ToBI's break index 2 might seem appropriate for such boundaries, it seems counter-intuitive to assign a lower index to boundaries that are perceived as stronger than intermediate boundaries.

model tends to blur the distinction, at least for those intermediate phrases that end in places within an utterance where a full-fledged boundary tone might also naturally have occurred. Indeed, such intermediate phrases may be considered to be full intonational phrases whose boundary tones have very little, if any, effect.[6] Consider the following question/answer pair, where the answer is uttered without hesitation and annotated with the notation described above.

(23) Q: I know that LISZT composed only ONE piano sonata. But how many did CHOPIN compose?

A: CHOPIN composed THREE piano sonatas.

L+H* L H* LL$

While careful analysis of the fundamental frequency contour might show a rise on the tail end of "composed," it is unclear whether the rise can be attributed to a high boundary tone (**H%**) ending the phrase "Chopin composed," or to effects of the H* accent which follows immediately on "three." If the answer is given with no hesitation between "composed" and "three," Pierrehumbert's **H%** does not seem to be warranted. Note, however, that in some contexts a pause may naturally occur between these two words. Consider a game show scenario where the host reveals the correct answer with a suspenseful pause, as demonstrated below.[7] In this context, an **H%** boundary between "composed" and "three" is almost certainly warranted.

[6] This generalization is perhaps applicable only to the **L** phrasal tone. Pierrehumbert and Hirschberg (1990, p. 302) note that an "**L** phrase accent emphasizes the separation of the current phrase from a subsequent phrase," whereas an "**H** phrase accent ... indicates the the current phrase is to be taken as forming part of a larger composite interpretive unit with the following phrase."

[7] In this example, the host is posing a "trick" question. Although Chopin did indeed write three piano sonatas, the first was written as a study at a very early age. Only the second (famous for the funeral march in the third movement) and the third are widely performed and considered representative of the mature Chopin.

(24) Host: Franz LISZT composed only ONE piano sonata.
How many did CHOPIN write?

Contestant: TWO.

Host: I'm sorry.

CHOPIN composed... *long suspenseful pause* ...
L+H* LH%
THREE piano sonatas.
H* LL$

While Pierrehumbert was primarily concerned with the manner in which phrasal tones affect the fundamental frequency contour at the end of intermediate phrases, the preceding examples illustrate that phrasal tones can also, in some cases, serve to delineate meaningful segments of an utterance in much the same way as boundary tones. Moreover, the fact that the phrase "Chopin composed" seems to demand a high boundary tone in example (24) implies that the choice of boundary tone is not random and may indeed reflect the discourse relationship between the two phrases (as we shall demonstrate later in the chapter). Since the same discourse relationship holds between the corresponding phrases in the non-hesitation case, it is reasonable to hypothesize that a pitch rise is intended after the phrasal low tone following "composed" regardless of whether or not there is a noticeable delay before the onset of the next intonational phrase. That is, just as intended phonemic segments are often lost when phrases such as "did you know" are contracted into "d'ya know," the effects of an intended rising boundary after "composed" in example (23) may be masked by the natural rise to the **H*** pitch accent in the subsequent phrase. Such situations are marked in our notation by placing the "intended" boundary in parentheses. For example, the first phrase in an utterance annotated with the tune **L+H* L(H%) H*LL$** is realized with pitch rise from end of the phrase to the lexical item bearing the **H*** accent with no hesitation. This notation encodes the fact that the first phrase, if ended with a slight pause, would in fact exhibit an **H%**

boundary. Moreover, it allows for a uniform treatment of full intonational phrases and intermediate phrases at the information structure level, a notion that will prove useful for the development of intonational grammars (as shown in Chapter 5). [8]

Using the notational conventions described above, each phrase boundary can be described by a combination of a phrasal tone and a boundary tone. As a result, example (23) can be rewritten as shown in example (25).

(25) CHOPIN composed THREE piano sonatas.
L+H* L(H%) H* LL$

Another important feature of this notation (which is also true of Pierrehumbert's notation and the ToBI conventions) is that it can specify an intonational contour without regard for the part of the utterance to which it applies. Thus, a given contour will have the same basic shape whether it covers a single word or a complete utterance. Similarly, contours can have a similar shape, but differ in their pitch range (i.e., the intensity of the f_0 peaks and valleys). Different intonational "tunes" composed of these elements are used to convey various discourse-related distinctions of "focus", given vs. new information, contrastiveness and propositional attitude (Grosz and Sidner 1986; Hirschberg 1990; Liberman and Pierrehumbert 1984; Ward and Hirschberg 1985; Wang and Hirschberg 1991; Prevost and Steedman 1994b; Prevost 1995). That is, they serve to indicate the status of the current phrase with respect to the surrounding phrases in the discourse.

3.3 Compositional Intonation

As the previous section suggested, a single intonational contour may be interpreted in a variety of ways, depending on the discourse context. Nevertheless, contours which share certain features (such as the same type of accent or the same boundary tone) often share certain aspects of meaning (Ladd 1980, Pierrehumbert and Hirschberg 1990). For instance,

[8] The assertion that some intermediate phrases (in Pierrehumbert's terminology) seem to behave in terms of information structural roles much like complete intonational phrases, should not be taken as an assertion that the distinction between intermediate and intonational phrases is never warranted. However, since the prosodic theory expounded here is concerned only with those intonational phrases corresponding to information structural delineations, described in Section 3.5, the remainder of this chapter will treat intermediate and intonational phrases and their information structural counterparts in a uniform manner.

pitch accents are used to identify certain information exchanged between the speaker and the listener as being particularly salient. Boundaries, on the other hand, which scope over an entire intonational phrase, often indicate the relationship between the current phrase and its successor. A high boundary tone (as in **LH%**) frequently implies that the speaker wishes to continue speaking or that the meaning of the current phrase should be interpreted with respect to its successor. Conversely, a low boundary tone (as in **LL%**) often denotes the end of a meaningful segment of speech. Not surprisingly, questions often end with a high boundary tone, leaving the topic under discussion open, while the sense of finality of a statement is generally marked by a low boundary.

The *meaning* of an intonation contour for a given utterance encompasses the relationships between the semantic content of the utterance and the beliefs of the discourse participants. That is, the meaning of a particular tune may communicate the speaker's beliefs about the proposition she is putting forth as well her beliefs about how that proposition will be interpreted by hearers. Furthermore, the choice of tune often conveys the speaker's beliefs about the nature of the relationships between propositions put forth in consecutive utterances or phrases. Considering the complex role of belief systems in determining intonational meaning, it is not surprising that to date no theory of intonational semantics has claimed to cover the wide range of prosodic variation in English. The remainder of this section, however, briefly discusses a recent compositional approach to intonational meaning and relates it to the generation task at hand.

Just as intonation contours are composed of pitch accents, phrasal tones and boundary tones in the physical sense, recent research has suggested that the meanings of intonation contours can be viewed as a composition of the meanings associated with each of the intonational parts (Pierrehumbert and Hirschberg 1990). This theory differs from the previous approaches that analyzed intonational meaning at the level of complete phrases or contours (e.g. Ladd 1980), and from those that attempted to define mappings between tunes and speaker attitudes (Jackendoff 1972, Ladd 1980, Ward and Hirschberg 1985). The principal advantage of the compositional model is that it captures that fact that tunes which share particular features often convey similar meanings. Such a compositional approach is also well suited for integration into a compositional semantics framework.

Just as the semantics of a particular string of words is composed of the semantics of its constituent parts, the corresponding intonational semantics can be derived from the intonational semantics of its parts.

Given the compositional nature of intonational representations, it will prove beneficial to examine the meanings of the intonational constituents of the tunes on which the present generation system relies most heavily. Without prematurely invoking too much detail, the tunes most widely employed in simple declaratory statements are {**H***}$^+$ **L**{**(L%), L%, L$, H%**} and {**L+H***}$^+$ **L**{**(H%),H%,H$**}, where the brackets are to be interpreted as selecting one element from the given set, and the "+" diacritic as standing for one or more occurrences of the symbol on which it occurs. Consequently, the components of primary interest are the **H*** and **L+H*** pitch accents and the **L(L%)**, **LL%**, **LL$**, **L(H%)**, **LH%** and **LH$** boundaries.

3.3.1 The Meaning of Pitch Accents

The **H*** accent is generally applied to items that are particularly salient. Although Pierrehumbert and Hirschberg (1990) and others have associated **H*** with items that are "new" to the discourse, it is certainly the case that items bearing an **H*** accent can be well established in the discourse if they require emphasis for some other reason, such as standing in contrast to another salient item. For example, given the question in (26), the occurrence of "pale" in the answer cannot be new information. Rather, it contrasts with the other available alternative, namely "brown." Regardless of whether an item marked by **H*** is new or contrastive, it is often utilized to instantiate the variable in the open proposition conveyed by the rest of the utterance (Prince 1986), and update the hearer's beliefs accordingly (Pierrehumbert and Hirschberg 1990).

(26) Q: Would you prefer the BROWN ale or the PALE ale?

L* H H* LL$

A: I'd like the PALE ale.

H* LL$

Although than the **H*** accent sometimes marks contrastive items, as shown above, the **L+H*** accent is associated with contrast more often. Since **H*** accents are so strongly

associated with (instantiated) variables in open propositions, **L+H*** accents are particularly useful for marking contrast in the fully instantiated background part of the open proposition, as noted by Jackendoff (1972) and others, and demonstrated in (27).

(27) Q: What about the beans? Who ate them?
A: Fred ate the beans
H*L L+H* LH$

3.3.2 The Meaning of Boundaries

The **L** phrasal tone has the effect of delimiting a segment that is to be interpreted separately from subsequent segments (Pierrehumbert and Hirschberg 1990). Recall that it is for this reason that the present approach disregards the semantic distinctions between intermediate and intonational phrases. If the **H** phrasal tone were to be included in the primary set of tunes handled by the present research, the intermediate/intonational phrase distinction would have to be maintained.

Boundary tones are often considered to mark the relationship between two consecutive phrases. An **H{%,$}** boundary tone frequently implies that the speaker wishes to continue speaking or that the meaning of the current phrase should be interpreted with respect to its successor. On the other hand, an **L{%,$}** boundary tone often denotes the end of a meaningful segment of speech that can be ostensibly interpreted on its own. While this interpretation seems correct in many cases, it does not adequately account for situations such as (27) where the open proposition carrying the **L+H* LH$** tune is to be interpreted with respect to the preceding phrase. Nonetheless, we maintain the belief that a high boundary marks information as somehow being incomplete.

3.4 Contrastive Stress Patterns

Due to the inability to easily derive semantic information from unrestricted text, text-to-speech systems are generally forced to rely on crude syntactic analyses and word classifications in making judgements about the accentability of words in an utterance, often using the strategy of *previous mention*, whereby a word is de-accented if it (or perhaps its root)

has previously occurred in some restricted segment of the text (Hirschberg 1990,Monaghan 1991). As described in the previous chapter, the text is often divided into such meaningful discourse segments on the basis of cue phrases and paragraph boundaries.

Meaning-to-speech systems differ from text-to-speech systems in the manner in which semantic and pragmatic information is exploited for assigning intonational features. Such systems have been employed in applications with limited, well-defined domains where semantic and discourse level knowledge is readily available. For these systems, the effectiveness of the previous mention strategy has been improved by considering semantic givenness in addition to lexical givenness when deciding if a word should be de-accented 1988.

Such enhanced previous-mention heuristics, while proving quite effective in practice, have exhibited several deficiencies that have been noted by their proponents. Foremost among these is the inability of such strategies to model the seemingly contrastive nature of many accentual patterns in spoken language (Hirschberg 1990). In some cases, contrastive stress errors may sound unnatural and in the worst case may actually mislead the hearer. Another problem that has been attributed to previous-mention strategies is the tendency to include too many accents (Monaghan 1991), potentially resulting in an inability for the hearer to determine the most important aspects of the speaker's intended message. The remainder of this section addresses these two problems and proposes explicitly modeling contrast in meaning-to-speech systems as a potential solution.

A previous-mention strategy might work as follows:

- Assign accents to open-class items (e.g. nouns, verbs, other content words)
- Do not assign accents to closed-class items (e.g. function words)
- De-accent any words that were already mentioned in the local discourse segment.

Now consider an application that allows users to configure a high-end stereo system and produces spoken descriptions of the system and its components. For simplicity, assume that there are five types of components available: amplifiers, pre-amps, CD players, tuners and speakers. Moreover, assume that each component is available in an American-made model and a British-made model, but that pairings of components made in the same country are not necessarily preferred. Such a program might be expected to produce the

type of output shown in (28) when a user fails to include an essential component in an audio system configuration.

(28) a. Your system does not include an AMPLIFIER.
b. The BRITISH amplifier comes HIGHLY RECOMMENDED.

Using a previous-mention algorithm like the one above will produce the appropriate accentual pattern on the NP *the British amplifier* in (28)b, stressing only *British*, because *amplifier* is explicitly mentioned in the previous sentence.

Now suppose the user selects a configuration consisting of an amplifier, a pre-amp and a tuner, all manufactured in Great Britain. Example (29) illustrates a possible excerpt from a description produced by the program, where b, b′, b″ and b‴ are alternative rather than successive utterances.

(29) a. Your system includes an AMPLIFIER, a PRE-amp and a TUNER.
b. The British AMPLIFIER you selected is HIGHLY RATED.
b′. The BRITISH AMPLIFIER you selected is HIGHLY RATED.
b″. The BRITISH amplifier you selected is HIGHLY RATED.
b‴. The British amplifier you selected is HIGHLY RATED.

The four accentual possibilities for the NP *the British amplifier* in the second sentence are given in (29)b-b‴. Examples (29)b and b′ are both acceptable because they correctly accent the contrastive *amplifier*, distinguishing it from the other available types of components. Based on the contents of the first sentence, however, the previous-mention strategy would produce the accentual pattern illustrated in (29)b″, which is clearly inappropriate. In fact, such an intonation may cause the hearer to infer that one of the components in the configuration was not British. Finally, if one considers the terms *british* and *amplifier* to be given prior to the utterance because of their inclusion in the configuration selected by the user, the previous-mention strategy would attempt to de-accent both terms as in (29)b‴. Since the NP clearly requires some form of accentuation, alternative strategies are necessary in such a case. Other plausible previous-mention strategies exhibit similar problems for equally simple examples.

Many of the problems associated with the previous-mention strategy in meaning-to-speech systems can be rectified by explicitly modeling contrastive stress. For the example

above, the program initially "knows" that the user's configuration includes a British *amplifier*, a British *pre-amp* and a British *tuner*. Hence, the program can construct an explicit set of alternative stereo components from which accentual patterns can be determined. By noting that the alternatives differ not in their manufacturing location, but in the actual type of component, the program can easily decide to stress *amplifier* rather than *British*. The precise algorithm for contrastive stress assignment is described in detail in Chapter 4.

By explicitly modeling contrastive stress, the over-accentuation problem of the previous-mention strategy can also be avoided. Consider a stereo configuration with two amplifiers: an American solid-state amplifier and a British solid-state amplifier. In describing such a configuration, the system might produce the utterance shown in (30).[9]

(30) The American solid-state amplifier is too powerful.

Using the previous-mention strategy would lead to the following accentual pattern on the subject noun phrase if the amplifier had not been mentioned previously.

(31) the AMERICAN SOLID-state AMPLIFIER is too POWERFUL.

The contrastive stress algorithm briefly sketched above is able to recognize the crucial distinction between the *American* and *British* properties of the system's two amplifiers and assign stress accordingly, producing the output in (32). Of course, that is not to say that *solid-state* and *amplifier* cannot bear pitch accents. In Chapter 4 we describe an approach for modeling both contrastive stress and the types of stress that might normally occur on such items that are new to the discourse.

(32) The AMERICAN solid-state amplifier is too POWERFUL.

3.4.1 Contrastive Pronouns

The examples above demonstrate the significance of accentual patterns in assisting the hearer's correct identification of referential items. Moreover, the examples illustrate the

[9] A closely related issue is how the system decides which modifiers are necessary in the description (Reiter and Dale 1992).

inadequacy of the notion of "givenness" as the sole model for determining accentuation. Further evidence that the previous mention heuristic is not a sufficient foundation for assigning accents is provided by the presence of stressed pronouns, which occur often in natural speech and are almost always contextually given.

While the failure to appropriately stress modifiers in a referring expression is liable to sound awkward and lead to misinterpretation, the failure to appropriately stress pronouns can frequently affect the semantic content of the proposition being conveyed. Recall that in example (19), repeated below as (33), the accentuation of "he" alters the meaning of the elided verb phrase.[10]

(33) a. John$_i$ thought BILL$_j$ would SUCCEED, but he$_j$ DIDN'T.
b. *Bill didn't succeed.*
c. John$_i$ thought BILL$_j$ would SUCCEED, but HE$_j$ didn't.
d. *Bill didn't think that Bill would succeed.*

The examples in (34) and (35) exhibits similar semantic shifts when the pronouns are stressed.[11]

(34) a. Bill$_i$ thinks RALPH$_j$ is quite DIM, but he$_i$ thinks HE$_i$ is BRILLIANT.
Bill thinks Bill is brilliant.
b. Bill$_i$ thinks RALPH$_j$ is quite DIM, but HE$_j$ thinks he$_j$ is BRILLIANT.
Ralph thinks Ralph is brilliant.
c. Bill$_i$ thinks RALPH$_j$ is quite DIM, but HE$_j$ thinks HE$_i$ is BRILLIANT.
Ralph thinks Bill is brilliant.

(35) a. John$_i$ called Bill$_j$ a REPUBLICAN, and then he$_i$ INSULTED him$_j$.
John insulted Bill.
b. John$_i$ called Bill$_j$ a REPUBLICAN, and then HE$_j$ insulted HIM$_i$.
Bill insulted John.

[10] For recent work on the effect of pitch accenting on pronoun referent resolution, see Cahn (1995).
[11] The reader should also note that the lack of a pitch accent on (35)b logically entails the proposition that calling someone a Republican is insulting.

3.4.2 Contrastive Stress in Naturally Occurring Discourse

The contrived examples in Section 3.4.1 indicate that stressed pronouns may pose a serious problem to the *previous mention* strategy of assigning pitch accents. They fail, however, to address the issue of how common the stressed pronoun phenomenon really is. At first glance, it may seem most appropriate to address the frequency of occurrence issue by examining naturally occurring speech and isolating examples of contrastively stressed pronouns. The principal concern with such an approach is the subjective nature of the definition of contrastive stress, as discussed in the previous chapter. Rather than conducting such a broad experiment, a simpler approach involving only explicitly contrastive constructions is taken here. The purpose of the present experiment is to determine how often subject pronouns are stressed in explicitly contrastive utterances. The data for the experiment is restricted to include only utterances of the form "but he ..." because the occurrence of stressed "he" in these utterances is more readily attributed to contrast than in utterances that are not syntactically marked as conveying contrastive information. Consequently, a certain degree of subjectivity is removed from the experiment.

The data for the experiment was extracted from the Switchboard corpus, a collection of over 2000 digitized telephone conversations collected at Texas Instruments. Since the corpus is a general tool for studying numerous aspects of speech data, the subjects were aware of neither the nature of the present experiment nor the intonational theories espoused by this writer.

In total, 162 occurrences of "but he ..." were extracted from 1022 conversations in the Switchboard corpus. Of those, 33 exhibited some degree of accentuation on "he," as determined by a combination of subjective judgments (by the author) and pitch track analyses. Of the 33 occurrences of stressed "he," two were eliminated as examples of contrastive stress because the immediately preceding discourse (approximately ten utterances) did not support such an interpretation. In the remaining cases, an antecedent for the stressed pronoun was clearly established in previous utterances. The results of the experiment, presented in Table 3.2, show that 31 (19.14%) of the 162 subject pronouns in explicitly marked contrastive constructions of the form "but he ..." received some degree

Disc	# He	# But He	# Stressed	# Contrastive	% Contrastive
disc01	157	4	0	0	0%
disc02	212	8	0	0	0%
disc03	265	6	1	1	16.67%
disc04	325	12	2	2	16.67%
disc05	245	14	0	0	0%
disc06	298	10	1	1	10.00%
disc07	326	14	4	4	28.57%
disc08	249	6	1	0	0%
disc09	277	9	3	2	22.22%
disc10	276	13	4	4	30.77%
disc11	274	13	2	2	15.38%
disc12	334	10	3	3	30.00%
disc13	427	21	8	8	38.10%
disc14	315	10	1	1	10.00%
disc15	312	12	3	3	25.00%
Totals	4292	162	33	31	19.14%

Table 3.2: Contrastive occurrences of "but he ..." in Switchboard Corpus (phase 1)

of stress that can reasonably be attributed to their contrastive status. Although it is impossible to extrapolate the data to cases where contrastive constructions are not marked explicity by coordinators like "but" or "however," the data does provide clear evidence that pronouns, despite their status as "given," are eligible to receive stress.

An analysis of data used for this experiment reveals the distinct pattern for contrastive stress shown in (36), where c and c' are to be considered alternatives to one another.

(36) a. {A & B are introduced}
b. A verb C,
c. But StressedPro(B) Neg(verb) C.
c′. But StressPro(B) Neg(do).

Examples of naturally occurring dialogues from the Switchboard corpus which follow this pattern are given in Examples 3.1 and 3.2 below.[12]

[12]For the sake of simplicity, these examples only explicitly show the stress on "he." The reader should not conclude that other words do not also bear noticeable pitch accents.

Example 3.1

A1 My, my little boy has gotten so into it that he's identified the, the people that have written certain songs, then he buys the pieces that have that person.

B1 Yeah.

A2 You know, on it.

B2 Oh, I see what you're saying.

A3 Huh?

B3 Yeah.

A4 I mean, I don't even know who did which ones, but HE does.

Example 3.2

A1 No, sometimes Mark and I, that's my husband Mark,

B1 # Uh-huh. #

A2 # um, # go to the August Moon which is down there it's a Chinese.

B2 # Oh, yeah, my husband likes that. #

A3 # uh, restaurant with the lions in # the front,

B3 yeah, off of Arapaho down there?

A4 yeah, and,

B4 Yeah, August Moon, uh-huh.

A5 Oh, it's pretty good as far as Chinese, but Chinese isn't my favorite so,

B5 I don't like it at all usually so [laughter].

A6 But HE likes it kind of ...

In each of these cases there are two directly contrastive elements and each contrastive entity is explicitly mentioned in the prior discourse. In other cases, the non-pronominal contrasts (i.e. the contrasts in the propositions ranging over the prominal) are somewhat less direct and require a bit more knowledge about the world. In Example 3.3, the contrast between hitting a ball straight down the middle and hitting it long requires some knowledge of the game of golf.

Example 3.3

A1 Well, we speak of the three games of golf here.

B1 Oh.

A2 And that would be the long, the short and the putt of it.

B2 Yes.

A3 So.

B3 The short of it is my weakest part.

A4 Uh-huh. We often times have made a trip to Mississippi in March because, of course, it's still cold here in March and we always hope for warm weather down there, and the people that we play with, the men, foursome, are all much longer ball hitters than my husband is. But HE finds that by going straight down the middle he usually wins about a quarter a hole because they've been in the rough on the right and then in the rough on the left. So HE ends up playing just as well as they do.

Furthermore, these examples do not contrast the stressed pronoun with only one other individual. Rather, they contrast the individual represented by the pronoun with a well-defined, salient set of individuals. Consequently, the statement "HE was sane," in Example 3.4 might be paraphrased as "Whitman, in contrast to other criminals who were insane, was sane."

Example 3.4

B1 Uh, you know, af-, even with the waiting period. And I, you know, I cannot see that, uh, that there is anybody that, that does not have criminal intent that would have, uh, any reason to object to that waiting period. That is,

A1 Uh-huh. That is right. And a lot of times like, uh, what is his name? I cannot remember his name right. The guy in the Austin tower, uh, Whitman?

B2 Yeah. Whitman.

A2 You know, HE was sane, and, and I guess they could have said like Lee Harvey Os-, Oswald was sane when he bought HIS gun, you know. But he bought HIS by mail anyway.

The data above clearly demonstrate the inadequacy of the notion of "givenness" to predict accent placement. Stressed subject pronouns in syntactically marked contrastive constructions are shown to occur with a frequency (19.14%) that cannot be overlooked when developing theories of accent placement and discourse context. While further work is necessary to determine the frequency of stressed pronouns in other types of sentences, the preliminary data indicate that the model of accent placement espoused here is better suited for handling contrastive accentual patterns than the *previous mention* heuristics employed by others.

The issues of contrastive stress are addressed further from a computational standpoint in Chapter 4 and Chapter 6. The latter chapter illustrates how the interaction of information structure, discourse structure and the contrastive pattern shown in (36) can be employed to appropriately stress contrastive items in automatically generated speech.

3.5 Information Structure

Information Structure refers to the organization of information within an utterance. In particular, information structure defines how the information conveyed by a sentence is related to the knowledge of the interlocutors and the structure of their discourse. Thus, sentences conveying the same propositional content in different contexts need not share

the same information structure. In simpler terms, information structure refers to how the semantic content of an utterance is packaged, and amounts to instructions for updating the information models of the discourse participants (Vallduví 1990). The realization of information structure in a sentence, however, differs from language to language. In English, for example, intonation carries much of the burden of information structure, while languages with freer word order, such as Catalan (Vallduví 1990) and Turkish (Hoffman 1995) convey information structure syntactically.

For the purposes of the present research, it will be convenient to view information structure as a bridge between intonation and discourse. This bridge connects the realization of intonational contours to high-level discourse structures that possess scant information about the target language or the method of communicating contextual information in the language. Consequently, information structure forms the competence model for the *semantics of discourse context*. The remainder of this section contains an examination of the notion of information structure in several different frameworks and using several different sets of terminology. The section concludes with a discussion of the relationship between models of discourse coherence and information packaging.

3.5.1 Information Structure Formalisms

Although a great number of linguists, computational linguists and phonologists have espoused theories of information structure, the lack of consistent terminology among competing models is a formidable obstacle to the understanding of the theories and their relationships to one another. Regardless of the diverse nomenclature, each of the frameworks for information structure share certain similarities. At the highest level, these theories all propose at least a two-way segmentation of utterances, where one segment represents a cognitive link to the discourse or the relevant knowledge pool, and the other represents the contribution of the sentence to the discourse or knowledge pool. This level of segmentation is intertwined with the given/new distinction, and indeed the two notions are often confused. Given the expositions in the previous sections about the accentuation of "given" discourse items, it should not be surprising that the present approach treats the given/new issue separately from the higher level information structural segmentation

issue.

Before delving into the confusing and intermingled vocabulary of information structure, it will be convenient to introduce a simple example, shown in (37).

(37) Q: As for the CRAZY chef, what did SHE cook?
A: (The CRAZY chef cooked)$_{th}$ (a rack of SPAM.)$_{rh}$

In this example, the material in parentheses with the subscript *th* corresponds to the link to prior utterances, or a pointer to a file card in Heim's framework (Heim 1983; also cf. Engdahl and Vallduví 1994, Hoffman 1995). This file card represents the discourse entity that is "updated" by the utterance. Conversely, the material marked by the *rh* subscript corresponds to the contribution that the speaker makes to hearer's model (i.e. the information that is to be stored on the "current" file card). The capitalized words, "crazy" and "spam," mark the location of pitch accents or *phonological focus*. This latter term should not be confused with broad interpretation of the term "focus" used by some information structure nomenclatures to describe the *rh* material in (37).

Table 3.3 presents the top-level terms which have been applied to the two segments in (37) by selected researchers. While the list is not exhaustive, it covers the core proposals from the information structure literature.[13]

The Vallduví (1990) system differs from the others in that it allows a three-way segmentation at the top level, including the *link*, *focus* and *tail*. The tail, which is not present in the previous example, includes background information that occurs after the focus. Such a representation is useful for cases such as (38), in which the background information in the question ("think will win") is split into non-contiguous parts in the answer.

(38) Q: Which team does Scott think will win the Stanley Cup next year?
A: (Of course he thinks) (the FLYERS) (will win).

Regardless of whether one assumes a two-way or three-way distinction among information structural components, this top-level segmentation alone cannot account for the

[13]The entry for Halliday is slightly misleading. In Halliday (1970), "information structure" is taken to refer to the given/new distinction, and the segmentation illustrated in (37) is termed "thematic" structure.

	The CRAZY chef cooked	a rack of SPAM
Halliday (1967,1970) Lyons (1977) Bolinger (1989) Steedman (1991a,1991c)	Theme	Rheme
Vallduví (1990) Engdahl and Vallduví (1994)	Link	Focus
Kuno (1976) Reinhart (1981) Hoffman (1995)	Topic	Comment
Hajičová and Sgall (1987,1988)	Topic	Focus
Prince (1986)	Open Proposition	Focus
Jackendoff (1972)	Presupposition	Focus
Grosz and Sidner (1986)	Focus	

Table 3.3: Information Structure Nomenclature

accentual pattern exhibited by (37), where elements of both segments receive pitch accents. Note that the given/new distinction is equally ill-suited to account for this stress pattern since both given information ("crazy") and new information ("spam") receive accents. In order to handle such intonational phenomena and the syntactic demands of free-word order languages, a deeper level of information structure more akin to phonological focus is required. The formalism advocated in the present research accomplishes this by specifying an additional level of information structure, as described in the following section.

3.5.2 An Intonational Approach to Information Structure

Because the concept of *focus* in the prosody literature is so frequently associated with phonological focus, and the terms *topic* and *focus* are both employed in the discourse literature, I have chosen to follow Steedman and Halliday in using the terms *theme* and *rheme* to describe the top-level information structural categories illustrated in example (37). Note, however, that Halliday's assumption that themes occur sentence-initially will be discarded, based on the types of situations found in Examples 3.5 and 3.6.

Example 3.5

Q: I know the AMERICAN amplifier produces MUDDY treble,

(But WHAT kind of treble) (does the BRITISH amplifier produce?)

L+H* L(H%) H* LL$

A: (The BRITISH amplifier produces) (CLEAN treble.)

L+H* L(H%) H* LL$

theme-focus-c *rheme-focus-c*

Theme *Rheme*

Example 3.6

Q: I know the AMERICAN amplifier produces MUDDY treble,

(But WHICH amplifier) (produces CLEAN treble?)

L+H* L(H%) H* LL$

A: (The BRITISH amplifier) (produces CLEAN treble.)

H* L(L%) L+H* LH$

rheme-focus-c *theme-focus-c*

Rheme *Theme*

Steedman (1991a) and Prevost and Steedman (1994b) argue that for the class of utterances exemplified by these examples, the rheme of the utterance occurs with an intonational phrase bearing a simple rise-fall tune. Using the variant of Pierrehumbert's abstract intonational notation described in Section 3.2, we write the rheme tune as **H* L(L%)**, **H* LL%**, or **H* LL$**. The theme of the utterance, in the cases when it does in fact include accented elements, bears the rise-fall-rise tune, written as **L+H* L(H%)**, **L+H* LH%** or **L+H* LH$**. [14]

[14] Recall that we are ignoring Pierrehumbert's distinction between intermediate and intonational phrase boundaries. The difference between the intonational boundaries L(L%), LL% and LL$ lies in the degree of lengthening and pausing (and similarly for L(H%), LH% and LH$). In our notation, the % symbol represents an intra-utterance boundary (as for a comma), and the $ symbol represents an utterance-final boundary. This distinction is often left implicit in the literature, both being written with % boundaries. For purposes of speech synthesis, however, the distinction is important since utterance boundaries must be accompanied by a greater degree of lengthening and pausing. Note also that the effect of the phrasal low tone preceding each of these boundaries propagates its effect back to the most recent pitch accent in the phrase.

Given the mappings between information structural constituents (themes and rhemes) and the intonational tunes described above, it should be clear that one might reasonably be able to determine appropriate tunes and boundaries when one is given information about how a sentence relates to its surrounding discourse. Note however, that simply knowing the string of words to which a tune applies does not imply that one knows the proper words on which to place the accents within the tune. That is, we need to address the issue of why the **H*** accent in 3.5 falls on "clean" rather than "treble." One possibility is that "clean" is new information in the answer, while "treble" was previously mentioned in the question. Another possibility, as discussed in Section 3.4, is that it is the property of being clean that distinguishes the treble mentioned in the answer from the (muddy) treble mentioned in the question.

As the examples demonstrate, both themes and rhemes can exhibit a combination of background and narrowly (phonologically) focused material. Hence, we differentiate instances of *theme-focus*, which mark contrastive items in themes, from instances of *rheme-focus*, which mark contrastive or new items in rhemes. Additionally, we will distinguish contrastive cases of theme- and rheme-focus by appending a "c" marker to the end of the category name, as shown in Examples 3.5 and 3.6. This differentiation between *theme/rheme-focus* and *theme/rheme-focus-c* is consistent with Rochemont's (1986) distinction between *presentational* and *contrastive* focus. Further support for such a distinction is provided by Vilkuna's (1989) account of contrastive foci in Finnish. Note also that nothing in the information structure representation prevents the occurrence of multiple theme- or rheme-foci(-c) within a top-level theme or rheme.

3.6 Models of Discourse Coherence

While the information structure representation adopted in the previous section encodes a level of semantic specificity necessary to account for intonational phrasing and accentual patterns, the present section addresses the problem of linking information structure to the discourse model. Of particular concern in this matter is the issue of how utterances are organized in meaningful ways, a notion often referred to as *discourse coherence* (Grosz,

Joshi and Weinstein 1986). The coherence of discourse can be analyzed with respect to both global structure, which defines how discourse segments comprising multiple utterances relate to one another, and local structure, which defines how consecutive utterances within such a segment relate to each other. The discussion below contains brief summaries of some of the well-known frameworks for discourse coherence and a discussion of the relevance of those models to information structure.

The *global* structure of discourse refers to the manner in which segments of the discourse are arranged with respect to its overall purpose. Grosz and Sidner (1986) refer to one aspect of this global organization as the *intentional* structure of the discourse. In this model, each discourse segment is associated with a discourse segment purpose (DSP) that contributes to the general discourse purpose (DP). Pierrehumbert and Hirschberg (1990) speculate that a hierarchical relationship between DSPs may be signaled by **H%** boundary tones, as exemplified by constructions where one utterance elaborates on the immediately prior utterance. While such observations are perhaps indicative of a relationship between top-level information structural divisions and DSP hierarchies, the relationship, if any, to thematic and rhematic focal articulations is unclear and warrants further research. Consequently, no claims are made in the present research regarding global discourse structure.

In addition to the global intentional structure, Grosz and Sidner (1986) also postulate a model of the *attentional* state of discourse, which represents the interlocutors' focus of attention. In the attentional state model, each discourse segment is associated with a local *focus space* which contains salient discourse entities for the given segment. These salient entities include items that have been explicitly mentioned as well as items that have been made salient by inferential reasoning. Unfortunately, this use of the term "focus" to refer to discourse-old entities is at odds with the information structural notion of focus as discourse-new or contrastive information. Consequently, for the purposes of the present research, the Grosz and Sidner terminology is avoided. Nonetheless, a similar structure called the *discourse entity list* is employed by the implementation described in Chapter 6 for the purpose of constructing referential expressions and resolving issues of givenness.

While the Grosz and Sidner model of attentional state provides a useful framework for

tracking discourse entities within a discourse segment, we appeal to the centering theory model (Grosz *et al.* 1986, Joshi and Weinstein 1981, Brennan *et al.* 1987) to describe local coherence relationships between consecutive utterances. Centering Theory is a formalism that traces entities through a discourse segment and describes the relationships between consecutive utterances. Though originally formulated as a predictive theory of discourse coherence, recent work has demonstrated the theory's usefulness for resolving pronoun ambiguities and describing constraints on the interpretation of *zero topics* in Japanese (Walker, Iida and Cote 1994).

Centering theory is based on a set of features associated with each utterance in a discourse segment, including the backward-looking center, a set of forward-looking centers, and a preferred center. The *backward-looking center* (C_b) is a discourse entity that links the current utterance with the previous utterance. The *forward-looking centers list* (C_f) contains the set of entities realized in the current utterance which are ordered by their possibly language specific potential to be the "center" of the subsequent utterance. The most highly ranked center in the C_f is called the *preferred center*, C_p.

The coherence of two consecutive utterances U_{i-1} and U_i is defined by a set of constraints, rules and transition states. The primary constraint states that for an utterance U_i, $C_b(U_i)$ is the most highly ranked item on the list $C_f(U_i)$ that occurs in the previous utterance, U_{i-1}. Pronominalization is directed by a rule that dictates that $C_b(U_i)$ must be realized as pronoun in U_i if any member of $C_f(U_i)$ is realized as a pronoun. Transition states, which define the coherence relationships between consecutive utterances, are ordered according to preference as shown in (39) and defined as shown in Table 3.4.[15]

(39) Continue > Retain > Smooth-Shift > Rough-Shift

The notion of *theme* described in Section 3.5 shares many similarities with the notion of a backward-looking center. Like a C_b, a theme is generally something that has already been established in the discourse and links the current utterance to previous utterances. Unlike a C_b, however, which is restricted to being a discourse entity, a theme may be a proposition over discourse entities. Consequently, the strongest possible claim one can

[15]This nomenclature for transition states is drawn from Walker *et al.* (1994).

	$C_b(U_i) = C_b(U_{i-1})$	$C_b(U_i) \neq C_b(U_{i-1})$
$C_b(U_i) = C_p(U_i)$	Continue	Smooth-Shift
$C_b(U_i) \neq C_p(U_i)$	Retain	Rough-Shift

Table 3.4: Centering Theory Transition States

make regarding the relationship between themes and C_bs is that the C_b of an utterance is realized within the theme of the utterance. Consider the simple example below, where "John" is the C_b and is undeniably part of the theme.

(40) a. When John heard about the lottery,

b. he immediately called his friend Bill.

c. (He told him)$_{th}$ (the good news right away.)$_{rh}$

Given that themes and C_bs are both defined in terms of links to prior utterances, the hypothesis put forth above certainly seems reasonable. Example (41), however, provides evidence that the C_b may occasionally occur in the rheme.

(41) U_1: John insulted BILL.

U_2: (Then BILL insulted)$_{th}$ (HIM)$_{rh}$.

L+H* L(H%) H* LL$

While the division of theme and rheme is this example is likely to be a matter of debate, the **H* LL$** tune on "him" is clearly the main phonological focus of the utterance (i.e the nuclear stress). In Vallduví's terminology, the file card corresponding to the discourse entity *Bill* is what is updated in U_2. The pronominalization of *John* in U_2, however, dictates that John must be the C_b. Although this preceding example complicates the relationship between theme-hood and C_b-hood, the discrepancy can be attributed to the fact that the C_f lists for the two utterances are identical and the rheme contains old information. When the rheme contains only new information, the C_b, if it exists, must be realized in the theme.

While themes and C_bs are clearly not identical structures, the fact that they both create links to prior utterances is a striking similarity. Consequently, the centering notion that some transitions between utterances are preferred over others can be adopted

without abandoning the information structure representations established in the previous section. We therefore consider the transition between two utterances that share thematic material to be akin to the continue and retain transitions of centering theory. Likewise, when the rhematic material of an utterance is realized in the theme of the following utterance, we conclude that something similar to centering's shift transitions has occurred. While centering theory and information structure are not strictly congruent, their similarities warrant the use of information structure as a model of discourse coherence. This perspective forms the basis for the organization of utterances produced by the program described in Chapter 6.

3.7 Summary

This chapter presented an *information structure* formalism that mediates between intonation and discourse, and encodes the proper level of semantic information to account for both contextually-bound accentuation patterns and intonational phrasing. The analysis of the link between intonation and discourse was divided into the following three components:

- **Intonational Meaning**: Intonation affects the meanings and interpretations of utterances. Just as the physical aspects of intonational contours are defined by the physical aspects of their components, so too are the meanings of intonational contours derived from the meanings of their components. Furthermore, intonation can reflect semantic contrasts among discourse entities, which are often signaled by the placement of pitch accents. A study of contrastive pronouns in naturally occurring speech provides further evidence for this assertion.

- **Information Structure**: The information structure of an utterance refers to the organization of information within the utterance and defines how the semantic material is related to the knowledge of the interlocutors and the structure of their discourse. A number of different nomenclatures for information structural divisions have been proposed. Due to the conflation of the information structural and phonological definitions of "focus," this terminology is supplanted by the terms

"theme" and "rheme." In order to avoid confusion, we adopt the terms theme-focus and rheme-focus to represent the phonological notion of focus within themes and rhemes. A correspondence between information structure and intonational tunes permits a semantic level of representation that encodes the contextual aspects of discourse that account for many types of intonational phenomena.

- **Discourse Coherence**: The notion of discourse coherence inherent in centering theory shares some striking similarities with the information structure notion of "linking" utterances. Unlike backward-looking centers, however, themes are not limited to being discourse entities, but instead may be propositions (from the discourse model) over entities in the discourse model. While centering theory and information structure are not strictly congruent, their similarities warrant the use of information structure as a model of discourse coherence.

The congruence between intonation and information structure and their similarities to the proposed discourse model are exploited by the Combinatory Categorial Grammar formalism presented in Chapter 5. As we shall see, these relationships also provide a framework for computational approaches to the spoken language generation problem.

Chapter 4

A Discourse-Based Semantic Focus Assignment Algorithm

The task of assigning an appropriate intonation contour to an utterance can be divided into two subtasks: determining the intonational phrasing (i.e. where boundaries occur), and determining the placement and shape of pitch accents within those phrases. As we saw in the previous chapter, intonational phrasing is associated with the theme/rheme articulation in the information structural representation. Similarly, the placement of accents within intonational phrases is associated with the focal articulations within themes and rhemes. This chapter describes the algorithms that produce the information structure representations for utterances. Section 4.1 describes how the centering theory model of coherence is applied to the problem of identifying the thematic structure of a proposition to be conveyed by the speaker. The algorithm for contrastive stress, which is based on sets of alternative discourse properties and entities, is presented in Section 4.2. The intonation assignment algorithms presented here are employed by a program that generates natural language descriptions of objects from a knowledge base, as described in Chapter 6.

Recall the familiar examples in (42) and (43) below.

(42) I know which amplifier produces clean BASS,

but (WHICH amplifier)$_{\text{th}}$ (produces clean TREBLE?)$_{\text{rh}}$
L+H* L(H%) H* LL$
(The BRITISH amplifier)$_{\text{rh}}$ (produces clean TREBLE.)$_{\text{th}}$
H* L(L%) L+H* LH$

(43) I know the British speaker produces MUDDY treble,

but (what kind of treble)$_{\text{th}}$ (does the British AMPLIFIER produce)$_{\text{rh}}$?
L+H* L(H%) H* LL$
(The British AMPLIFIER produces)$_{\text{th}}$ (CLEAN treble.)$_{\text{rh}}$
L+H* L(H%) H* LL$

As noted above, the answers to these two *wh*-questions contain the exact same string of words, but very different theme/rheme divisions and intonational patterns. Moreover, note that the information structural bracketing in (43) does not correspond to the standard right branching grammatical bracketing (i.e. $S \rightarrow NP\ VP$). Because the semantic chunking of thematic and rhematic material does not always correspond to traditional notions of syntactic constituency, the algorithm presented below relies on a grammatical formalism called Combinatory Categorial Grammar (CCG, Steedman 1991a) to realize the intonational decisions it makes in the (phonological and syntactic) surface form. CCG, which is described in detail in Chapter 5 allows a more flexible notion of constituency than other more traditional grammars in order account for prosodic phrasing as well as certain syntactic phenomena such as non-constituent coordination. Without giving the details of the grammar, the algorithm for determining appropriate information structural bracketings and assigning (phonological) foci within themes and rhemes is described below.

4.1 Prosodic Phrase Identification

Given the mapping between information structural constituents and intonational tunes described in Section 3.5, we now turn our attention to the problem of determining how to divide an utterance into its theme and rheme. From the examples above it should be clear that in making such determinations, the discourse context must be explicitly modeled. Several hierarchical discourse models (Grosz and Sidner 1986; Polanyi 1988) have been proposed which involve segmenting the discourse into discrete chunks based on a variety of segmental criteria (e.g. attentional structure, intentional structure). These theories, which describe how discourse segments relate to one another, are concerned with *global* discourse coherence. Other theories of discourse, such as Centering Theory (Grosz, Joshi and Weinstein 1986), are more concerned with the *local* coherence *within* a discourse segment. For our purposes, we will loosely define a discourse segment as a paragraph primarily about a single object or concept, and will concern ourselves with the local coherence among sentences within such segments.

Centering Theory is a formalism that tracks entities through a discourse segment and describes the relationships between consecutive utterances. For each utterance, the theory identifies a backward-looking center (C_b), a discourse entity that links the current utterance with the previous utterance. In addition, the theory posits a list of discourse entities, called the forward looking centers list (C_f), which are ordered by their (possibly language specific) potential to be the "center" of the subsequent utterance. For an utterance U_i, $C_b(U_i)$ is the most highly ranked item on the list $C_f(U_i)$ that occurs in the previous utterance, U_{i-1}. The most highly ranked center in the $C_f(U_i)$ is called the *preferred center*, $C_p(U_i)$. The relation between $C_b(U_i)$ and $C_b(U_{i-1})$, and the relation between $C_b(U_i)$ and $C_p(U_i)$, determines the type of transition between utterances. Of the four possible transitions, only the *continue* and *retain* transitions, which occur when $C_b(U_i) = C_b(U_{i-1})$, will be of primary concern, for it is these transitions that apply to text tightly coupled to a description of a single discourse entity.

The notion of *theme* described in Section 3.5 shares many similarities with the notion of a backward-looking center. Like a C_b, a theme is something that has already been established in the discourse. Unlike a C_b, which is restricted to being a discourse entity,

a theme, as defined above, may be a proposition over discourse entities.

Following Centering Theory, a coherent division of an utterance into its theme and rheme can be determined by examining themes and rhemes that have already been established in the discourse. Consequently, the discourse model consists of a collection of themes and rhemes derived from previous utterances and ordered by recency, which we call the information structure store (ISstore). Because it is necessary to compare themes and rhemes to one another, entries in the ISstore correspond to abstract semantic representations of themes and rhemes rather than the many possible strings of words that may realize them. Indeed, throughout the remainder of this paper, the terms theme and rheme will refer to such representations. Given a semantic representation for a complete utterance, its theme can be determined by lambda-abstracting over variables and searching the ISstore top-down until a unifying theme or rheme is found. Based on the Centering Theory model of local coherence, the most recent match is accepted, giving a preference to thematic matches, which by definition result in the preferred continue or retain transitions. Having determined the theme of the utterance, its rheme is determined by abstracting over the theme in the full semantic representation. That is, given the complete semantics of the utterance, and the semantics of its thematic argument, one can abstract over that argument to determine the semantics of its rheme.[1]

Consider a scenario in which a natural language generator has the intention of producing a spoken description of an entity from a knowledge base, say the "x4" amplifier. To bootstrap the discourse processing, assume that the sole entry in the ISstore is $\{\texttt{theme}{:}\lambda P.(Px4')\}$. Suppose we wish to determine the division into theme and rheme of the semantic representation shown in (44)a, based on the prior discourse as represented by the ISstore. By comparing (44)a against the ISstore, we see that the theme on the top of the store and the semantic representation in (44)b can β-reduce to form (44)a. Consequently (44)b represents the rheme of the utterance in question. We can then update the ISstore to include the new rheme as well the repeated theme. Moreover, by abstracting over the variables in the rheme, we can add several other propositions to the ISstore, allowing future utterances to thematicize particular elements of the rheme, as allowed by

[1] Of course, the semantic representation of the theme may itself be a lambda function.

Centering's smooth-shift and rough-shift transitions.

(44) a. $mass(x, (treble'(x) \& clean'(x)) \& produce'(x4', x))$

b. $\lambda y.mass(x, (treble'(x) \& clean'(x)) \& produce'(y, x))$

c. (The x4)$_{\text{th}}$ (produces clean treble.)$_{\text{rh}}$

4.2 Pitch Accent Specification

Having established a framework for identifying the appropriate top-level information structural divisions, the problem of identifying the proper placement of theme- and rheme-foci within those IS constituents can now be addressed. In the previous section we explained at length how a *previous mention* strategy can be used to predict pitch accents in text-to-speech systems. The system described here adopts that basic strategy, but also includes an algorithm for assigning *contrastive focus* which serves to distinguish among multiple salient discourse entities (Prevost 1995).

The previous mention strategy is invoked as follows. For each predicate in the semantic representation scheme, the system keeps track of whether or not it has been evoked in the discourse. Thus, when processing a semantic representation like the one in (44)b, the system merely needs to check a few boolean flags to determine that `produce'`, `clean'` and `treble'` are new to the discourse.[2] The system considers the concepts represented by these symbols to be in phonological focus (henceforth focus) and marks them in the semantic representation with the $\circ$ predicate, as shown in (45)a. The consequence of making this move is the eventual generation of the accentual pattern in (45)b.

(45) a. λy.mass(x, ($\circ treble'(x) \& \circ clean'(x)) \& \circ produce'(y, x)$)

b. (The x4)$_{\text{th}}$ (PRODUCES CLEAN TREBLE.)$_{\text{rh}}$

H* H* H* LL\$

As demonstrated in Section 3.4, the previous mention strategy often fails to accentuate items that must be stressed in order to distinguish among competing discourse

[2]Because of the limitations imposed on the types of monologues the system attempts to generate and their relatively short lengths, we ignore the problem of determining how long it takes for an evoked discourse entity to fade from the discourse model.

items. To address this problem, we introduce a second form of phonological focus, called *contrastive* focus, and mark items requiring such focus with the • predicate in semantic representations. For example, in the context of the question in (46)a, the system produces the representation for the rheme shown in (46)b, which eventually leads to the accentual pattern shown in (46)c.

(46) a. Which kind of treble does the x4 produce?

b. $\lambda P.mass(x, (treble'(x) \& \bullet\, clean'(x)) \& (Px))$

c. (The x4 produces)$_{\text{th}}$ (CLEAN treble.)$_{\text{rh}}$

H*$_c$ LL\$

Based on an instrumental analysis of spoken data from the Switchboard corpus using tools such as *xwaves*, slightly more prominent pitch accents (marked H*$_c$) are assigned to items bearing contrastive focus. In the eventual synthesized output of the system, these accents have higher pitch targets than the H* accents. The presence of different levels of H* pitch accents is left implicit in the Pierrehumbert notation system, but clearly must be modeled explicitly in a system that *generates* intonation.

The basic premise on which this model of contrastive focus assignment stands is that each object in the database is associated with a set of *alternatives* from which it must be distinguished if reference is to succeed. When the high-level generator constructs a referring expression which includes a number of properties, those properties in the expression that are *not* common to all elements of the set of alternatives are marked with the contrastive focus operator •.

The data structures and notational conventions given in (47) are useful for describing the focusing algorithm.

(47) *DElist*: a collection of discourse entities that have been evoked in prior discourse, ordered by recency. The list may be limited to some size k so that only the k most recent discourse entities pushed onto the list are retrievable.

ASet(x): the set of alternatives for object x, i.e. those objects that belong to the same class as x, as defined in the knowledge base.

RSet(x, S): the set of alternatives for object x as restricted by the referring expressions in *DElist* and the set of properties S.

CSet(x, S): the subset of properties of S that must be accented for contrastive purposes.

Props(x): a list of properties for object x, ordered by the grammar so that nominal properties (e.g. *amplifier*) take precedence over adjectival properties (e.g. *British*).

The algorithm, which assigns contrastive focus in both thematic and rhematic constituents, begins by isolating the discourse entities in the given constituent. For each such entity x, the structures defined above are initialized as follows:

(48)
$$\begin{aligned} \textit{Props}(x) &:= [P \mid P(x) \text{ is true in the database }] \\ \textit{ASet}(x) &:= \{y \mid \textit{alt}(x, y)\}, \text{ the set of alternatives for } x \\ \textit{RSet}(x, \{\}) &:= \{x\} \cup \{y \mid y \in \textit{ASet}(x) \;\&\; y \in \textit{DElist}\} \\ \textit{CSet}(x, \{\}) &:= \{\} \end{aligned}$$

The algorithm appears in pseudo-code below:

(49) $S := \{\}$

for each P **in** $Props(x)$

$RSet(x, S \cup \{P\}) := \{y \mid y \in RSet(x, S)\ \&\ P(y)\}$

if $RSet(x, S \cup \{P\}) = RSet(x, S)$ **then**

% no restrictions were made based on property P.

$CSet(x, S \cup \{P\}) := CSet(x, S)$

else

% property P eliminated some members of the *RSet*.

$CSet(x, S \cup \{P\}) := CSet(x, S) \cup \{P\}$

endif

$S := S \cup \{P\}$

endfor

In other words, given an object x, a list of its properties and a set of alternatives, the set of alternatives is restricted by including in the initial *RSet* only those objects that are explicitly referred to in the prior discourse and x itself. Initially, the set of properties to be contrasted (*CSet*) is empty. Then, for each property of x in turn, the *RSet* is restricted to include only those objects satisfying the given property in the knowledge base. If imposing this restriction on the *RSet* for a given property decreases the cardinality of the *RSet*, then the property serves to distinguish x from other salient alternatives evoked in the prior discourse, and is therefore added to the contrast set. Conversely, if imposing the restriction on the *RSet* for a given property does not change the *RSet*, the property is *not* necessary for distinguishing x from its alternatives, and is not added to the *CSet*.

The algorithm can be most easily understood by working through an example. Consider the simple Prolog representation in (50) below and suppose that all of the entities in the set {`c1, c2, c3, c4`} are alternatives to one another.

(50)
```
amplifier(c1).
amplifier(c2).
tuner(c3).
tuner(c4).
british(c1).
british(c3).
american(c2).
american(c4).
produce(c1,clean_treble).
```

Now consider the question in example (51), and suppose that the discourse entity in the rheme of the response is *c1*, the British amplifier.

(51) Q: Does the *British* amplifier or the *American* amplifier produce clean treble?

The successive states of the algorithm are illustrated in Example 4.1. After initializing the various data structures, an attempt is made to restrict the elements of the *RSet* based on the first property on the property list (*amplifier*). Since both stereo components mentioned in the question are amplifiers, the *RSet* is left unchanged and there is no need to add the *amplifier* property to the contrast set. Next, the algorithm considers the *British* property, altering the *RSet* to exclude those procedures that do not share the property of being manufactured in Great Britain. Since the restriction *does* change the *RSet*, the *British* property is added to the contrast set. Having reduced the *RSet* to one element, the algorithm produces the rheme in Example (52) and eventually the answer in (53), accenting only the linguistic material in the rheme corresponding to the properties in the *CSet*.

(52) Rheme: $\lambda P.def(x, (\bullet british'(x) \& amplifier'(x)) \& (Px))$

(53) A: (The BRITISH amplifier) (produces clean treble).

Now consider Example 4.2, which produces a rather different accentual pattern. In this example the *British* property remains unstressed in the response. This is due to the

Example 4.1

Question	
Does the BRITISH amplifier or the AMERICAN amplifier produce clean treble?	
Initial State	
$ASet(c1)$	$\{c1, c2, c3, c4\}$
$Props(c1)$	$[amplifier, british]$
$RSet(c1, \{\})$	$\{c1, c2\}$
$CSet(c1, \{\})$	$\{\}$
After *Amplifier* Property Restriction	
$RSet(c1, \{amplifier\})$	$\{c1, c2\}$
$CSet(c1, \{amplifier\})$	$\{\}$
After *British* Property Restriction	
$RSet(c1, \{amplifier, british\})$	$\{c1\}$
$CSet(c1, \{amplifier, british\})$	$\{british\}$
Resulting Rheme	
$\lambda P.def(x, (\bullet british'(x) \& amplifier'(x)) \& (Px))$	
Result	
(The BRITISH amplifier) (produces clean treble.)	

fact that restricting on this property does not serve to distinguish any subset of the *RSet* from the other elements of the set.

Example 4.3 illustrates the case where a modifier (*British*) is unstressed even though a contrastive modifier (*American*) is realized in the question. The reason for this, as previously stated, is because *British* and *American* are restricting over different types of objects in the discourse model (i.e. *amplifiers* vs. *tuners*). Again, the fact that one can reasonably leave *British* unstressed in the response does not imply that it must remain de-accented. A slight modification to the focusing algorithm as presented above allows this second possibility. We merely allow the propositions to make their restrictions based not only on the *RSet* from the most recent iteration, but also on the (possibly) larger *RSet* from the previous iteration. Thus, in example 4.3 we can perform the *British* property restriction based on $RSet(c1, \{\})$ as well as $RSet(c1, \{amplifier\})$, thereby allowing *British* to enter the contrast set. Note that this modification does not adversely affect the

Example 4.2

Question	
Does the British AMPLIFIER or the British TUNER produce clean treble?	
Initial State	
$ASet(c1)$	$\{c1, c2, c3, c4\}$
$Props(c1)$	$[amplifier, british]$
$RSet(c1, \{\})$	$\{c1, c3\}$
$CSet(c1, \{\})$	$\{\}$
After *Amplifier* Property Restriction	
$RSet(c1, \{amplifier\})$	$\{c1\}$
$CSet(c1, \{amplifier\})$	$\{amplifier\}$
After *British* Property Restriction	
$RSet(c1, \{amplifier, british\})$	$\{c1\}$
$CSet(c1, \{amplifier, british\})$	$\{amplifier\}$
Resulting Rheme	
$\lambda P.def(x, (british'(x) \& \bullet\ amplifier'(x)) \& (Px))$	
Result	
(The British AMPLIFIER) (produces clean treble.)	

results of the simpler examples shown above. Furthermore, the algorithm remains able to produce the proper results for examples which require multiple accents, as shown in Example 4.4.

The examples presented above illustrate how a simple version of alternative set semantics can be used to determine accent assignments for conveying contrastive distinctions within referring expressions.

4.3 Summary

This chapter presented algorithms for processing discourse information and building the semantic and information structural representations necessary for generating context-appropriate intonation. The top-level division of an utterance into theme and rheme is based on a collection of themes and rhemes from prior utterances called the ISStore. The theme of an utterance is determined in a manner similar to the determination of the C_b

Example 4.3

Question	
Does the BRITISH AMPLIFIER or the AMERICAN TUNER produce clean treble?	
Initial State	
ASet(c1)	{c1,c2,c3,c4}
Props(c1)	[*amplifier*,*british*]
RSet(c1,{})	{c1,c4}
CSet(c1,{})	{}
After *Amplifier* Property Restriction	
RSet(c1,{*amplifier*})	{c1}
CSet(c1,{*amplifier*})	{*amplifier*}
After *British* Property Restriction	
RSet(c1,{*amplifier*,*british*})	{c1}
CSet(c1,{*amplifier*,*british*})	{*amplifier*}
Resulting Rheme	
$\lambda P.def(x,(british'(x)\& \bullet amplifier'(x))\&(Px))$	
Result	
(The british AMPLIFIER) (produces clean treble.)	

in centering theory.

The lower-level articulations of theme-focus and rheme-focus are produced by a two phase algorithm. The first phase invokes a *previous mention heuristic* similar to those proposed in earlier related work. The second phase, which accounts for instances of contrast, is based on sets of alternative properties and entities that are salient in the discourse. The algorithm works by successively considering properties of the given entity with respect for their ability to distinguish that entity from the other alternatives. The chapter provides several examples of the algorithm's effectiveness.

Example 4.4

Question	
Does the BRITISH AMPLIFIER, the AMERICAN AMPLIFIER or the BRITISH TUNER produce clean treble?	
Initial State	
ASet(c1)	{c1,c2,c3,c4}
Props(c1)	[*amplifier*,*british*]
RSet(c1,{})	{c1,c2,c3}
CSet(c1,{})	{}
After *Amplifier* Property Restriction	
RSet(c1,{*amplifier*})	{c1,c2}
CSet(c1,{*amplifier*})	{*amplifier*}
After *British* Property Restriction	
RSet(c1,{*amplifier*,*british*})	{c1}
CSet(c1,{*amplifier*,*british*})	{*amplifier*,*british*}
Resulting Rheme	
$\lambda P.def(x,(\bullet british'(x)\& \bullet amplifier'(x))\&(Px))$	
Result	
(The BRITISH AMPLIFIER) (produces clean treble.)	

Chapter 5

Intonation and Combinatory Categorial Grammar

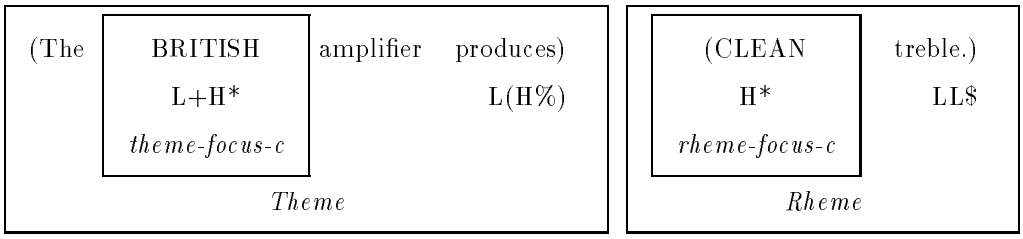
While the semantic, prosodic, and discourse-level structures introduced in the previous chapters form the basis of a competence model for organizing the propositions to be put forth in spoken language, it is the job of the syntactic competence model to dictate the mapping between such propositions and strings of spoken words. Note, however, that the higher-level semantic, prosodic and discourse models which have been proposed for natural language generation often impose structures and constraints that are orthogonal to those imposed by traditional notions of syntactic constituency. For example, consider the question/answer pair in Example 5.1 below, where the theme of the answer includes the subject NP as well as the verb, thereby crossing the traditional subject/predicate bracketing.

Example 5.1

Q: I know the AMERICAN amplifier produces MUDDY treble,

(But WHAT kind of treble) (does the BRITISH amplifier produce?)

L+H* L(H%) H* LL$



Many "traditional" grammatical formalisms (e.g. transformational grammar) license syntactic bracketings orthogonal to a given prosodic structure of an utterance and forbid certain syntactic bracketings that are isomorphic to the prosodic structure (e.g. ⟨subject verb⟩ ⟨object⟩). If we accept such a competence model for language, it becomes necessary to introduce separate prosodic and syntactic components into any computational model that aims to simulate spoken language production. The inclusion of separate processes which operate on entirely disparate structures necessarily introduces computational complexity and obfuscates the issue of whether semantic interpretations should be assigned to syntactic constituents or to intonational phrases.

It is for these reasons that we adopt the more flexible notion of syntactic constituency offered by Combinatory Categorial Grammar (CCG), a formalism that has been advocated for handling constructions involving non-traditional syntactic constituents such as those shown in (54).

(54) a. John admired, but Mary detested, the newest member of the team.

b. Mary detested the member of the team who John admired.

c. John gave the dog a bone, and the policeman a flower.

Moreover, it has been noted by Steedman (1990b,1991a,1991c,1991b) and others (Moortgat 1989, Oehrle 1988) that the types of constituents required to account for prosodic phrasing appear to coincide with those required by the preceding examples, as well as with other types of constituents that have motivated the development of such grammars.

This chapter presents Combinatory Categorial Grammars for governing syntax and intonation, and demonstrates how such grammars can be utilized in tandem to restrict the space of available parses for an utterance. Section 5.3 presents a simple algorithm for producing semantic and information structural representations from prosodically annotated utterances. An algorithm for the reverse process—generating sentences with intonational markings from focus-marked semantic representations of the message to be conveyed—is presented in Section 5.4.

5.1 Combinatory Categorial Grammars

Although context-free grammars (CFGs) are quite attractive from a computational standpoint, requiring only the power of pushdown automata, their usefulness for natural language processing is somewhat limited. In fact, natural languages are generally believed to fall outside of the class of context-free languages (CFLs) (Chomsky 1957, Postal 1964, Shieber 1985) based on certain constructions such as crossed-serial dependencies in Dutch. Consequently, many computational linguists have turned to *mildly context-sensitive grammars* which are computationally tractable (i.e. polynomially parsable), but also able to generate some languages outside of the class of CFLs (Joshi 1985). Combinatory Categorial Grammars (CCGs, Steedman 1990b,1991a,1991c,1991b), which are an extension of pure Categorial Grammars (Ajdukiewicz 1935, Bar Hillel 1953), are lexicalized grammars that generate such mildly context-sensitive languages.[1] A *lexicalized* grammar is one in which each lexical item is associated with a set of constraints that determines the ways in which it may combine with other lexical items or syntactic constituents. Such grammars are generally associated with a very small set of rules or operations that govern syntactic derivations. In contrast, non-lexicalized grammars for natural languages, such as CFGs, generally contain relatively impoverished lexical representations and include a very large number of rules governing syntactic derivations.

Although the choice of CCG for the syntactic component over more traditional grammars is fundamental for maintaining the ability to process (i.e. parse and generate) portions of utterances corresponding to syntactic, semantic and prosodic phrases in tandem,

[1] Pure CGs include only functional application rules and are weakly equivalent to CFGs.

it is not the only type of grammatical formalism that is adaptable to this property. For example, there are a variety of other categorial formalisms that produce similar types of non-traditional syntactic constituents (Moortgat 1989, Oehrle 1988, Dowty 1988). Moreover, there are non-categorial formalisms, such as Tree-Adjoining Grammar (TAG) (Joshi 1985) and Linear-Indexed Grammar (LIG) that have been shown to possess generative power that is weakly equivalent to CCG (Weir 1988), where weak equivalence is defined as follows.

Definition 5.1 *Two grammars, G and G' are* weakly equivalent *if $L(G)$, the set of strings generated by G, is identical to $L(G')$, the set of strings generated by G'. The underlying derivation structures of the strings of $L(G)$ and $L(G')$ need not be identical.*

5.1.1 A Formal Definition of CCGs

A combinatory categorial grammar is associated with a set of non-terminals, V_N, which are said to be *atomic*. Based on the atomic categories, the complete set of possible categories, $C(V_N)$ for a CCG, G, can be defined recursively as follows (Weir 1988):

Definition 5.2 *For a CCG, G, with atomic categories V_N,*

i. c is a category of G if $c \in V_N$, the set of atomic categories of G.

ii. c_r/c_a is a category of G if c_r and c_a are categories of G.

iii. $c_r\backslash c_a$ is a category of G if c_r and c_a are categories of G.

The non-atomic categories c_r/c_a and $c_r\backslash c_a$ are each said to be *functional* categories whose argument category is c_a and whose result category is c_r. Any functional category can be written using only the forward and backward slashes (/ and \ respectively), the set V_N of atomic categories and left and right parentheses. Since both types of slashes are left associative, a category of the form $((...(c_1|_1c_2)|_2...)|_{n-1}c_n)$ may be unambiguously written as $c_1|_1c_2|_2...|_{n-1}c_n$, where the vertical slash ($|_x$) is a variable ranging over the set $\{/, \backslash\}$.

Given the definition of its categories, a CCG can be defined as follows by the quintuple (V_T, V_N, s, f, C) (Weir 1988):

Definition 5.3 *Let G be a CCG. Then $G \stackrel{\text{def}}{=} (V_T, V_N, s, f, R)$, where*

V_T *is a finite set of lexical items (terminals in a derivation tree).*

V_N *is a finite set of atomic categories, a subset of the complete set of categories (nonterminals) in a derivation tree.*

s *is a distinguished element of V_N.*

f *is a function $V_T \rightarrow \mathcal{P}(C(V_N))$, which maps each element of V_T to a finite subset of the power set of $C(V_N)$ (i.e. to a finite set of categories).*

R *is a finite set of combinatory rules governing how categories may combine.*

The generative power of a CCG is dependent on its set R of combinatory rules. The most basic rules for a categorial grammar are the functional application rules, given below in Definitions 5.4 and 5.5. Note that the directionality of the slash in a CCG category dictates whether it can apply to another category on its left or its right.

Definition 5.4 *Two adjacent CCG categories, X/Y and Y, can combine by forward functional application as follows:*

$$X/Y \qquad Y \qquad \rightarrow \qquad X$$

where X and Y are variables over $C(V_N)$.

Definition 5.5 *Two adjacent CCG categories, Y and $X\backslash Y$, can combine by backward functional application as follows:*

$$Y \qquad X\backslash Y \qquad \rightarrow \qquad X$$

where X and Y are variables over $C(V_N)$.

Given these two combinatory rules alone, the CCG is capable of generating context-free languages such as $a^n b^n$, but unable to generate context-sensitive languages such as wcw, where $w \in \{a, b\}^+$. A simple CCG for generating the former is shown in Example 5.2

below. A derivation for the string *aaabbb* involving only the forward functional application rule is given in Example 5.3.

Example 5.2

A CCG for generating $a^n b^n, n \geq 1$. Let $G = (\{a, b\}, \{s, b'\}, s, f, R)$, where

$$f(a) = s/b', s/b'/s$$

$$f(b) = b'$$

and R contains the rules in Definitions 5.4 and 5.5.

Example 5.3

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{a}{s/b'/s} \quad \dfrac{a}{s/b'/s} \quad \dfrac{\dfrac{a}{s/b'} \quad \dfrac{b}{b'}}{s}}{s/b'} \quad \dfrac{b}{b'}}{s} \quad \dfrac{b}{b'}}{s/b'}}{s}$$

By introducing some other rules into the set R, the generative power of CCGs can be expanded beyond the context-free languages. For present purposes, we shall consider two other types of rules, functional composition rules and type raising rules. The former types of rules compose functional categories in the mathematical sense. That is, for functions f and g, a new function $f \circ g$ is created such that the result of f serves as the input to g. Although there are many types of functional composition that may be included in a CCG, we present only the generalized forward composition rule, which will prove useful for English derivations (for $n = 1$).

Definition 5.6 *Two adjacent CCG categories, X/Y and $Y|_1 Z_1|_2 \ldots|_n Z_n$, where $n \geq 1$, can combine by forward functional composition as follows:*

$$X/Y \qquad Y|_1 Z_1|_2 \ldots|_n Z_n \qquad \rightarrow \qquad X|_1 Z_1|_2 \ldots|_n Z_n$$

where X, Y and Z are variables over $C(V_N)$. At $n = 0$, this rule reduces to the forward application rule.

Note that the simple inclusion of the forward composition rule allows one to derive the string *aaabbb* from the grammar of Example 5.2 with a derivation structure different from Example 5.3, as illustrated below.

Example 5.4

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{a}{s/b'/s} \quad \dfrac{a}{s/b'/s}}{s/b'/b'/s} \quad \dfrac{a}{s/b'}}{s/b'/b'/b'} \quad \dfrac{b}{b'}}{s/b'/b'} \quad \dfrac{b}{b'}}{s/b'} \quad \dfrac{b}{b'}}{s}$$

Type raising rules allow a category to switch its function or argument role in a CCG derivation. For example, a category c, which may normally be consumed by functional application by a category of the form T/c, may raise its type to that of $T\backslash(T/c)$, i.e. a category that takes a functional category over c and returns the result of that functional category.

While the inclusion of composition and type raising rules expands the generative power of a given CCG, the usefulness of the CCG formalism stems from the ability to adequately restrict its power to generate only the subset of context-sensitive languages appropriate for a natural language such as English. This problem is addressed below.

5.1.2 CCGs and Natural Languages

Combinatory Categorial Grammars have been employed to describe syntactic phenomena in a number of natural languages. Steedman (1985) has shown their usefulness for a variety of English constructions, as well as Dutch subordinate clause crossed dependencies, as illustrated in (55).

(55) ...Jan Piet Marie zag helpen zwemman. (from Bresnan *et al.* 1982)
...Jan Piet Marie saw help swim.
"...Jan saw Piet help Marie swim."

More recently, Hoffman (1995) has presented an extension of CCG, called multiset-CCG, to handle free word order phenomena in Turkish.

For present purposes, we describe a simple CCG for English involving the set of atomic categories $\{s, np, n, pp\}$ and the functional application and composition rules described above. Thus, we write the syntactic category for the transitive verb *produce*, in its most simple form, as:

(56) produce $\equiv (s\backslash np)/np$

We can also include agreement features on syntactic categories by transforming atomic categories such as np into terms such as $np(3, s)$, where the first argument gives the agreement feature for *person* and the second argument gives the agreement feature for *number*. Consequently, the categories for the transitive verb *produce* become:

(57) produce $\equiv (s\backslash np(1, N_{subj}))/np(P_{obj}, N_{obj})$
produce $\equiv (s\backslash np(2, N_{subj}))/np(P_{obj}, N_{obj})$
produces $\equiv (s\backslash np(3, singular))/np(P_{obj}, N_{obj})$
produce $\equiv (s\backslash np(3, plural))/np(P_{obj}, N_{obj})$

In this example, and indeed throughout the remainder of this chapter, we follow the Prolog convention of identifying variables with capital letters. For the categories listed above, the variable P_{obj} ranges of the set $\{1, 2, 3\}$, while N_{subj} and N_{obj} range over the set $\{singular, plural\}$. Variables such as these are instantiated by invocations of the derivation rules described below.

Since it will be useful to build semantic representations for CCG derivations, it will also be convenient to assign semantic interpretations to each CCG category. One simple approach is to associate a term in the lambda calculus with each syntactic category. For example, the lambda expression corresponding to the categories in (57) might be:

(58) produce $\equiv \lambda Y.\lambda X.produce'(X,Y)$

Note that the ordering of the curried variables X and Y in the above expression is entirely dependent on constraints imposed by the corresponding syntactic category. That is, since the syntactic category for *produce* is principally a function over an object NP (i.e. an NP to its right), the semantic interpretation must be a function over the interpretation of an object NP. While a cursory comparison of the syntactic category and its semantic representation reveals the correspondences between its syntactic and semantic arguments, the notation can be somewhat simplified by directly associating the syntactic categories with semantic interpretations. This is accomplished by using an infix colon operator (:) which has higher precedence than the forward and backward slashes, but lower precedence than the syntactic and semantic terms it connects. Consequently, the categories for the verb *produce* can be written as shown below, where X and Y are variables over the semantic interpretations of noun phrases.

(59) produce $\equiv (s : produce'(X,Y)\backslash np(1, N_{subj}) : X)/np(P_{obj}, N_{obj}) : Y$
produce $\equiv (s : produce'(X,Y)\backslash np(2, N_{subj}) : X)/np(P_{obj}, N_{obj}) : Y$
produces $\equiv (s : produce'(X,Y)\backslash np(3, singular) : X)/np(P_{obj}, N_{obj}) : Y$
produce $\equiv (s : produce'(X,Y)\backslash np(3, plural) : X)/np(P_{obj}, N_{obj}) : Y$

Thus, the category for the verb *produces* is a function from a noun phrase (meaning Y) on its right, to another function from a third person, singular noun phrase (meaning X) on its left, to a sentence (meaning $produce'(X,Y)$).[2]

While the types of categories shown above clearly convey language-specific information about relative word orders, it is the derivation rules that license certain combinations of words while prohibiting others. Moreover, it is the derivation rules that define how semantic interpretations are combined. Since the categories over which the derivation rules operate may not be fully instantiated, as illustrated in (59), the rules necessarily impose certain unification constraints. In this manner, variables corresponding to syntactic agreement features and semantic interpretations are instantiated by the derivation.

[2]There are obviously a number of other features that CCG categories might encode, such as voice and tense for verbs. While these features are omitted from the present examples for the purpose of avoiding unnecessary complication, the interested reader is referred to Appendix A for numerous examples of how such features can be represented.

As discussed earlier, the forward and backward application rules are sufficient for performing context-free derivations. These rules say that when a CCG function is next to a category that unifies with its expected argument (on the proper side given by the direction of the slash), then the two can be combined and assigned the result category of the function. The application rules for English are shown below in Definitions 5.7 and 5.8.

Definition 5.7 *Two adjacent categories which unify with X/Y and Y respectively, can combine by forward functional application as follows:*

$$X/Y \qquad Y \qquad \rightarrow \qquad X$$

In derivations, forward application is abbreviated with the symbol ">".

Definition 5.8 *Two adjacent categories which unify with Y and $X\backslash Y$ respectively, can combine by backward functional application as follows:*

$$Y \qquad X\backslash Y \qquad \rightarrow \qquad X$$

In derivations, backward application is abbreviated with the symbol "<".

Given the two application rules, one can easily derive sentences with a traditional right branching structure, as shown in Example 5.5.[3]

Example 5.5

```
   Speakers                        produce                        sound
=================  ====================================  ==============
np(3,p):speakers'  (s:produce'(X,Y)\np(3,p):X)/np(P,N):Y  np(3,s):sound'
                   ------------------------------------------------------>
                                s:produce'(X,sound')\np(3,p):X
-------------------------------------------------------------<
              s:produce'(speakers',sound')
```

These rules, however, are not enough to provide the range of derivations required to account for prosodic structure and the flexible surface constituency required for the

[3]Note the effects of the rules' unification constraints in producing the semantic interpretation in the result category.

English constructions illustrated in (54). In order to handle such cases, the CCG for English must minimally include a forward composition rule and restricted type raising, extending the power of the formalism beyond the context-free fragment described above. The restricted ($n = 1$) forward composition rule, shown in Definition 5.9, allows two functions to combine if and only if the result category of the right function unifies with the argument category of the left function.

Definition 5.9 *Two adjacent categories which unify with X/Y and Y/Z respectively, can combine by forward functional composition as follows:*

$$X/Y \qquad Y/Z \qquad \rightarrow_{\mathbf{B}} \qquad X/Z$$

Forward composition is identified in derivations by the symbol ">**B**,*" reminiscent of Curry's composition combinator.*

We shall also stipulate that arguments of verbs must possess type-raised categories. Consequently, the category for the subject noun phrase "speakers" is a function from a predicate (i.e. a verb phrase) on its right to a sentence, as shown in (60). Two possible categories for the type-raised object noun phrase "sound" are given in (61) and (62). In (61), the category for "sound" is a function from a transitive verb on its left to a predicate (verb phrase), whereas in (62), the category for "sound" is a function from a sentence missing its object to a complete sentence.

(60) speakers $\equiv s : S/(s : S\backslash np(3,p) : speakers')$

(61) sound $\equiv (s : S\backslash np(P_{subj}, N_{subj}) : X)\backslash((s : S\backslash np(P_{subj}, N_{subj}) : X)/np(P_{obj}, N_{obj}) : sound')$

(62) sound $\equiv s : S\backslash(s : S/np : sound')$

As a result of including the forward composition and type-raising rules, the simple transitive sentence "*Speakers produce sound*" can be derived with both right branching and left branching structures, as shown in Examples 5.6 and 5.7.[4]

[4]For the sake of simplicity, the np categories are shown without their agreement features.

Example 5.6

```
      Speakers                    produce                          sound
=====================  ==========================  ================================
s:S/(s:S\np:speakers')  (s:produce'(X,Y)\np:X)/np:Y  (s:S\np:X)\((s:S\np:X)/np:sound')
                       ------------------------------------------------------------<
                                              s:produce'(X,sound')\np:X
-------------------------------------------------------------------->
                  s:produce'(speakers',sound')
```

Example 5.7

```
      Speakers                    produce                    sound
=====================  ==========================  ==================
s:S/(s:S\np:speakers')  (s:produce'(X,Y)\np:X)/np:Y  s:S\(s:S/np:sound')
-------------------------------------------------->B
           s:produce'(speakers',Y)/np:Y
           ------------------------------------------------------------<
                        s:produce'(speakers',sound')
```

The two derivations produce interpretations with identical function-argument structure, but quite different syntactic bracketings. It is precisely this flexible notion of constituency that permits syntactic bracketings and prosodic phrasing to be structurally isomorphic. For example, the bracketing imposed by 5.6 corresponds to the prosody one would expect if the sentence were uttered in response to the question "What produces sound?" as shown in (63). Conversely, the bracketing imposed by 5.7 corresponds to the prosody one would expect if the sentence were uttered in reponse to the question "What do speakers produce?" as shown in (64).

(63) (SPEAKERS) (produce SOUND.)
H* L(L%) L+H* LH$

(64) (SPEAKERS produce) (SOUND.)
L+H* L(H%) H* LL$

5.2 Prosody and CCG

The examples of the previous section serve to illustrate the flexible nature of syntactic constituency inherent in the CCG formalism. Moreover, the syntactic bracketings licensed by CCG in the preceding examples are all structurally isomorphic to some ostensibly "natural" prosodic phrasing for some appropriate context. Consequently, if a grammar for intonation can be defined, it should produce prosodic constituents that reflect the structure of a valid CCG syntactic bracketing. Then, rather than requiring two completely separate processes for handling syntax and prosody, as proposed by Selkirk (1984), the two "isomorphic" grammars can be employed in tandem to generate spoken sentences. The remainder of this section is devoted to defining such prosodic grammars and illustrating their usefulness in constraining syntactic derivations in CCG.

The basic strategy for invoking the syntactic and prosodic grammars in concert is given by the prosodic constituent condition in Definition 5.10 below. This condition stipulates that the two grammars must operate in lockstep. That is, two constituents may combine to yield a larger constituent if and only if they are licensed by the CCG rules to combine both syntactically *and* prosodically.

Definition 5.10 *The Prosodic Constituent Condition:*
Combination of two syntactic categories using a syntactic combinatory rule is only allowed if their corresponding prosodic categories can also combine by some (possibly different) prosodic combinatory rule. (Steedman 1990b,1991a,1991c,1991b)

Since the computational complexity of two grammars operating in tandem depends on their individual complexities, the generative capacity of the prosodic grammar cannot exceed that of the syntactic grammar without adversely affecting computational tractability. In fact, the prosodic CCGs presented below, which require only the forward and backward functional application rules and a rather severly limited composition rule, fall within the class of context-free grammars.

Although the Prosodic Constituent Condition defines how a prosodic grammar can be linked to a syntactic grammar in the CCG framework, it does not establish a methodology for constructing a prosodic grammar. The discussion below introduces such a methodology

by appealing to the similarities between prosodic and syntactic constraints.

5.2.1 Groundwork for Prosodic Grammars

When linguists build grammars to describe syntactic phenomena, they often model miniscule subsets of a natural language such as English. The idea is generally to illustrate the way a particular type of grammatical formalism can handle certain interesting syntactic constructions without necessarily involving the full complexity of the language being modeled. Likewise, computational linguists often build highly constrained grammars with the specific requirements of a computational application in mind. In presenting grammars for governing prosody, a similar approach is possible. Therefore, the prosodic grammars offered below are not intended to cover the wide range of intonational patterns in English, but rather to model the prosodic variations in a tightly constrained mode of discourse. The purpose of such an exercise is not to present steadfast, wide-coverage prosodic grammars for English, but to present a methodology for constructing prosodic grammars and employing them in tandem with syntactic grammars. Rather than constructing a single prosodic grammar to illustrate this point, three stages of a prosodic grammar's incremental development are described below.

Before delving into the process of describing prosodic grammars, however, it will be necessary to clarify some rather important distinctions between the framework set forth by Pierrehumbert (1980) and the one proffered here. Recall that Pierrehumbert defines an intonational phrase as a collection of intermediate phrases followed by a boundary tone, where an intermediate phrase is defined as a sequence of pitch accents followed by a phrasal tone. This approach was borne purely from the analysis of spoken data rather than any attempt at generating prosodic parameters for artificial speech. Although the Pierrehumbert notation minimizes the number of intonational diacritics for boundaries, it fails to make distinctions concerning the degree of lengthening or pausing associated with the boundaries. As a result, the notation provides no method of differentiating between sentence-final and intra-sentential boundaries, which are certainly not prosodically identical. Because the present research is aimed at prescribing a set of rules for governing prosody in synthetic speech, such distinctions are crucial.

Tune	Description
H* L(L%)	Utterance-initial complete rheme
H* LH%	Non-final partial rheme
H* LL$	Utterance-final rheme
L+H* L(H%)	Utterance-initial theme
L+H* LH$	Utterance-final theme

Table 5.1: Inventory of Tunes

Tune	Description
H^*_c L(L%)	Utterance-initial complete rheme
H^*_c LH%	Non-final partial rheme
H^*_c LL$	Utterance-final rheme
L+H^*_c L(H%)	Utterance-initial theme
L+H^*_c LH$	Utterance-final theme

Table 5.2: Inventory of Tunes (with contrastive accents)

Recall that a notation for handling the granularity of boundaries needed for speech synthesis was proposed in Section 3.2.3. Given this revised notation and the mapping between top-level information structural constituents and intonation contours established in Section 3.5, a subset of contours for generating *wh*-questions and simple declaratory statements can be easily defined. Table 5.1 shows the basic inventory of tunes which concern the present research.[5] If, for the purposes of generation, it is stipulated that lexical items bearing contrastive accents receive more emphatic stress, the tunes shown in Table 5.2 must also be admitted into the model.

Based on the notational conventions and mappings described above, the following properties form the basis for constructing prosodic grammars.

(i) A boundary must combine with at least one pitch accent to its left.

(ii) A boundary may not combine with another boundary.

(iii) Constituents which are prosodically unmarked may freely combine with non-boundary

[5]The correspondence of these tunes to the information structure constituents is based on the limited types of discourse under discussion. The reader should not infer that other tunes do not portray such functions in other situations.

constituents bearing prosodic information.

(iv) Multiple pitch accents may occur in an intonational phrase.

(v) A complete intonational phrase may combine only with another complete intonational phrase.

(vi) A constituent of any length bearing no pitch accents can promote itself to a full thematic intonational phrase. (Null Theme Promotion Rule)

While the first four of these assumptions are clearly derived from Pierrehumbert's system (as modified above), the fifth assumption relates to information structural aspects of prosodic grammar constituents. Recall that the discourse processing model presented in Chapter 3 postulates that information structural units (i.e themes and rhemes) update the discourse model as they are discovered. By stipulating that complete phrases (corresponding to complete themes or rhemes) can only combine with other complete phrases, the grammar is necessarily prohibited from assigning any constituent the category of a complete phrase unless there are no surrounding constituents that could potentially be part of the given phrase. Consequently, there can be no ambiguity about what point in a derivation the discourse model should be updated.

Because of the structural isomorphism among intonational structure, syntactic structure and information structure, prosodic CCGs based on the five assumptions listed above can operate in tandem with syntactic/semantic CCGs in accordance with the aforementioned prosodic constituent condition, which essentially synchronizes the grammars. In addition to this condition, however, there are two additional ways in which the grammars interact. First, the grammars associate the placement of pitch accents within an utterance with the focused elements of its semantic/information structural representation. This is accomplished in the lexicon by associating each lexical item with both focused and unfocused categories. For example, while the category for *produces* normally appears as shown in (65), it assumes the category in (66) when it bears a pitch accent marking new information, and assumes the category in (67) when it bears a pitch accent marking contrastive information.

(65) produces $\equiv (s : produce'(X,Y) \backslash np(3, singular) : X) / np(P,N) : Y$

(66) produces $\equiv (s : \circ produce'(X,Y) \backslash np(3, singular) : X) / np(P,N) : Y$

(67) produces $\equiv (s : \bullet produce'(X,Y) \backslash np(3, singular) : X) / np(P,N) : Y$

The second way in which the prosodic and syntactic grammars interact is governed by the categories assigned to intonational boundaries. Since the prosodic effects of boundaries may stretch over several words in an utterance, it seems inappropriate to associate them with single lexical items in the same manner that pitch accents correspond to lexical items. Rather, boundaries are associated with *linked* categories in the prosodic grammar. A linked boundary category provides both a syntactic/semantic category as well as a prosodic category, often with unification constraints between the two. The syntactic/semantic part of boundary category generally copies the syntactic/semantic category to its left (i.e. the category for the constituent comprising the intonational phrase) and marks it as a complete information structural theme or rheme appropriately, thereby capturing the intention behind assumption (v) above.

A simple approach to building a prosodic CCG grammar based on the aforementioned five assumptions is to assign pitch accents the category of a function over boundaries, as described in previous versions of the present research (Steedman 1991a,Prevost and Steedman 1993a). For example, the prosodic category for a constituent bearing the **H*** pitch accent might be given as $rheme/bl$. Note however that the combination of complete intonational phrases is dependent on the boundary types, as evidenced by the fact that sentence-initial and sentence-final themes and rhemes possess different boundary tones (or at least demand different break indices). In order to encode such information in the functional categories for pitch accents, it becomes necessary to create distinct result types for each possible type of boundary that a given pitch accent might precede. While it is certainly possible to include multifarious categories in the prosodic lexicon for each pitch accent, the fact that pitch accents occur more frequently in speech than boundaries implies that the derivational search space will be substantially larger than if the combinatory properties of complete phrases were actually encoded in the boundary categories themselves. Consequently, by raising the types of boundaries to be functions over the (functional) pitch accent categories, the boundary categories can appropriately encode

information regarding the combination of complete intonational phrases.

5.2.2 Building Prosodic CCGs

Among the intonational patterns that have been discussed in previous chapters, the simplest types of utterances include a single thematic phrase (marked by the **L+H* L(H%)** tune) followed by a single rhematic phrase (marked by the **H* LL$** tune), or a single rhematic phrase (marked by **H* L(L%)**) followed by a single thematic phrase (marked by **L+H* LH$**). Numerous examples have illustrated the usefulness of such contours for simple *wh*-questions and their responses. Example 5.8 offers an assignment of categories to such utterances and their constituent thematic and rhematic phrases.[6]

Example 5.8

Utterance Level Constituents:

{L+H*}$^+$ L(H%)	{H*}$^+$ LL$	:=	u
{H*}$^+$ L(L%)	{L+H*}$^+$ LH$	:=	u

Phrase (Information Structural) Level Constituents:

{H*}$^+$ LL$	:=	rh
{H*}$^+$ L(L%)	:=	u/th
{L+H*}$^+$ LH$	:=	th
{L+H*}$^+$ L(H%)	:=	u/rh

Grammar 5.1, which is described in detail below, illustrates the suitability of CCG to model the prosodic possibilities in Example 5.8.[7] While this grammar allows for instances of multiple phonological focus (i.e. multiple pitch accents) within a phrase, it fails to distinguish between contrastive and new focus. The grammar does however account for unaccented words which may occur before, after and between pitch accents in a phrase, as illustrated in Example 5.1 at the beginning of this chapter.

[6] As in Chapter 3, the "+" diacritic is used in the *regular expression* sense, requiring one or more occurrences of the item it marks.

[7] The variable T in these examples ranges over syntactic types. E, which ranges over the set $\{rh,th\}$ stands for "-emes."

Grammar 5.1

Pitch Accent Categories:

H*	:=	rh/bl,	$(rh/bl)/(rh/bl)$
L+H*	:=	th/bh,	$(th/bh)/(th/bh)$
Ø	:=	X/X,	where X is a variable
			and $\forall Y, Z \quad X \neq Y \setminus Z$

Boundary Categories:

LL$	:=	$T@E \setminus T$
		$E\backslash(E/bl)$
L(L%)	:=	$T@(u/th) \setminus T$
		$(u/th)\backslash(E/bl)$
LH$	:=	$T@E \setminus T$
		$E\backslash(E/bh)$
L(H%)	:=	$T@(u/rh) \setminus T$
		$(u/rh)\backslash(E/bh)$

There are several interesting aspects of this grammar that warrant discussion. First, there are two types of categories assigned to pitch accents: functions over boundaries and functions over other pitch accents of the same type. Consequently, two similar pitch accents may combine by one invocation of the functional application rule, and three similar pitch accents may combine either by two instances of functional application, or by one instance each of composition and application. Note however, that the composition rule required in the latter case need not be as general as the composition rule described above in Definition 5.9. In the prosodic case, the composition rule can be restricted to cases where the rightmost category has the same argument and result types, as captured by the definition of prosodic forward functional composition below.

Definition 5.11 *Two adjacent prosodic categories which unify with X/Y and Y/Y respectively, can combine by forward prosodic functional composition as follows:*

$$X/Y \qquad Y/Y \qquad \rightarrow \qquad X/Y$$

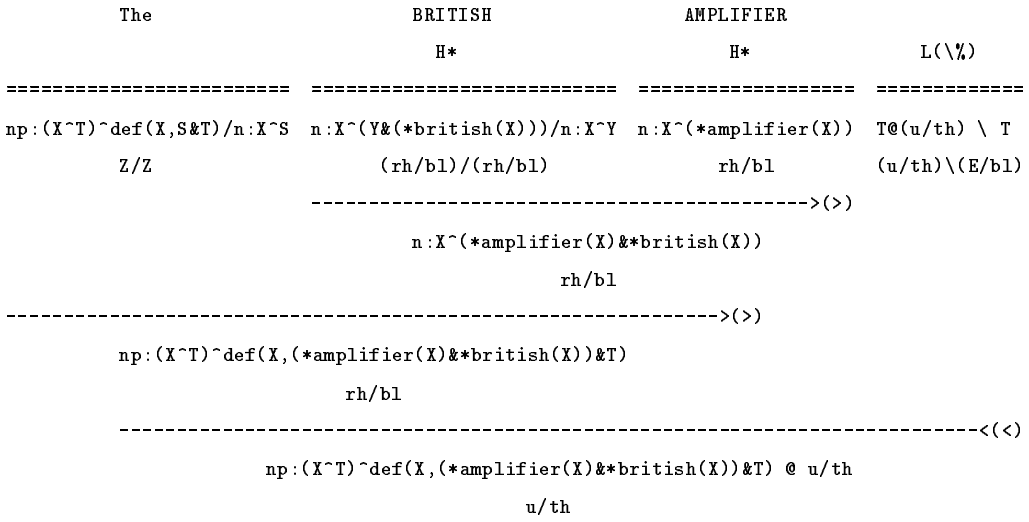
In derivations, forward composition is abbreviated by the symbol $>B$.

For lexical items that are not associated with a pitch accent, we assign the null tone category X/X, which can combine with any pitch accent category to its right by forward application, or with any pitch accent category to its left by forward prosodic composition, to yield a constituent bearing the same pitch accent category. Additionally, an item bearing the null tone category can combine with another such item to its right by prosodic functional composition. Consequently, the null tone category allows intonational tunes to be spread over arbitrarily long constituents such that only certain lexical items within a constituent are associated with pitch accents. The additional constraint on the null tone category X/X, that X cannot unify with any category that unifies with $Y\backslash Z$, ensures that lexical items bearing no pitch accent cannot combine with boundary categories to their right, thereby preserving the property that boundaries can only combine with an otherwise complete phrasal constituent.

Since boundaries are not associated with single lexical items, they are not assigned syntactic/semantic (henceforth *synsem*) categories by the syntactic grammar. Therefore, the boundary categories in the prosodic grammar (Grammar 5.1) include both synsem and prosodic parts. When a boundary combines with a boundary-less but otherwise complete phrase to its left, the synsem part of the boundary category combines by backward application with the synsem category to its left, yielding the same category annotated with a phrase-level prosodic category. That is, a synsem category representing a complete intonational (and hence information structural) phrase is associated in a derivation with its prosodic category by the infix @ operator, which takes higher precedence than the categorial slashes. This point is illustrated in Example 5.9.[8] Note that in derivations such as this, which involve both syntactic and prosodic combinatory rules, the symbol denoting the particular combinatory rule involved in a derivation step is written in parentheses immediately following the symbol denoting the syntactic rule.

[8] In this derivation, and those that follow, the $\circ$ operator is represented by an asterisk.

Example 5.9



Since categories of the form $X@Y$ cannot combine with any other categories under the application rules set forth above, an additional rule for reducing two such categories must be included in the grammar, as defined below.

Definition 5.12 *Two syntactic categories $X@A$ and $Y@B$* reduce *to category $Z@C$ if and only if there is some syntactic combinatory rule such that $X \quad Y \quad \rightarrow \quad Z$ is a valid derivational step and there is some prosodic combinatory rule such that $A \quad B \quad \rightarrow \quad C$ is a valid derivational step.*

The effect of this rule is that complete intonational (and hence information structural) phrases can only combine with other complete intonational phrases, as desired. For example, the phrasal-level categories shown in (68)a and (68)b can combine to form the utterance-level category in (68)c.[9]

[9](68)a represents the type raised result of the derivation in Example 5.9.

(68) a. $s : def(X, (\circ amplifier(X) \& \circ british(X)) \& T)/(s : T \backslash np : X)@u/th$

u/th

b. $s : mass(Y, (treble(Y) \& \circ clean(Y)) \& produces(X, Y)) \backslash np : X@th$

th

c. $s : def(X, (\circ amplifier(X) \& \circ british(X)) \& mass(Y, (treble(Y) \& \circ clean(Y))$
$\& produces(X, Y)))@u$

u

Of course, Grammar 5.1 covers a range of prosodic possibilities that is somewhat more limited than those examined in Chapter 3. Recall that those accounts of prosody and information structure allowed for multiple propositions, of which some more general proposition is an abstraction, to be realized in an utterance, thereby resulting in more than two intonational phrases. For example, the propositions in (69) can both be abstracted to the open proposition in (70), representing a type-raised entity. The open propositions in (71) are the complements, with respect to the β-reduction operation of the lambda calculus, of (70).

(69) a. *(∘graduated-from′ ∘penn′ john′)*

b. *(∘currently-attends′ ∘harvard′ john′)*

(70) $\lambda P.(P john')$

(71) a. *λ x. (∘graduated-from ∘penn′ x)*

b. *λ y. (∘currently-attends ∘harvard′ y)*

Taken together, the propositions in (70) and (71) might be realized in the sentence shown in (72). [10]

(72) (John) (GRADUATED from PENN),
H* H* LH%

(and CURRENTLY attends HARVARD).
H* H* LL$

[10] Although the accents (72) are contrastive in nature, they are marked as H* rather than H^*_c in order to simplify the example.

Suppose we wish to construct a grammar that covers such examples by allowing multiple rhematic phrases to combine together. We might assign categories to utterance-level and phrase-level constituents as shown in Example 5.10. This scheme allows any number of **H* LH%** phrases to appear to the left of an utterance final rheme, and any number of such phrases to combine together to form an utterance initial rheme. Examples of the types of utterances licensed under this scheme are shown in (73).

These prosodic constraints are captured in Grammar 5.2 by assigning appropriate categories to the **LH%** boundary.

(73) a. (JOHN) (HATES MOZART).
L+H* LH% H* H* LL$

b. (JOHN) (HATES MOZART) (but LOVES BEETHOVEN).
L+H* LH% H* H* LH% H* H* LL$

c. (JOHN HATES) (MOZART).
H* H* L(L%) L+H* LH$

d. (JOHN HATES) (but JOAN LOVES) (MOZART).
H* H* LH% H* H* LH% L+H* LH$

Example 5.10

Utterance Level Constituents:

$\{\text{L+H*}\}^+\ \text{L(H\%)}\quad \{\text{H*}\}^+\ \text{LL\$} \quad := \quad u$

$\{\text{L+H*}\}^+\ \text{L(H\%)}\quad \{\{\text{H*}\}^+\ \text{LH\%}\}^+\quad \{\text{H*}\}^+\ \text{LL\$} \quad := \quad u$

$\{\text{H*}\}^+\ \text{L(L\%)}\quad \{\text{L+H*}\}^+\ \text{LH\$} \quad := \quad u$

$\{\{\text{H*}\}^+\ \text{LH\%}\}^+\quad \{\text{L+H*}\}^+\ \text{LH\$} \quad := \quad u$

Phrase (Information Structural) Level Constituents:

$\{\text{H*}\}^+\ \text{LL\$} \quad := \quad rh$

$\{\text{H*}\}^+\ \text{L(L\%)} \quad := \quad u/th$

$\{\text{H*}\}^+\ \text{LH\%} \quad := \quad rh/rh,\quad u/th_2/(u/th_2),\quad u/th\backslash(u/th_2/(u/th_2))$

$\{\text{L+H*}\}^+\ \text{LH\$} \quad := \quad th$

$\{\text{L+H*}\}^+\ \text{L(H\%)} \quad := \quad u/rh$

Grammar 5.2

Pitch Accent Categories:

H*	:=	$rh/bl,$	$(rh/bl)/(rh/bl)$
L+H*	:=	$th/bh,$	$(th/bh)/(th/bh)$
Ø	:=	$X/X,$	where X is a variable and
			$\forall Y, Z \quad X \neq Y\backslash Z$

Boundary Categories:

LL$	:=	$T@E \setminus T$
		$E\backslash(E/bl)$
L(L%)	:=	$T@(u/th) \setminus T$
		$(u/th)\backslash(E/bl)$
LH%	:=	$T@(E/E) \setminus T$
		$(E/E)\backslash(E/bl)$
	:=	$T@(u/th_2/(u/th_2)) \setminus T$
		$u/th_2/(u/th_2)\backslash(E/bl)$
	:=	$T@(u/th)\backslash(u/th_2/(u/th_2)) \setminus T$
		$(u/th)\backslash(u/th_2/(u/th_2))\backslash(E/bl)$
LH$	:=	$T@E \setminus T$
		$E\backslash(E/bh)$
L(H%)	:=	$T@(u/rh) \setminus T$
		$(u/rh)\backslash(E/bh)$

Example 5.11 illustrates the derivation for the first phrase in the utterance "the BRITISH amplifier, which is HIGHLY RATED, is the most EXPENSIVE," intoned as if in response to the question "which amplifier is most expensive?"

Example 5.11

```
          The                        BRITISH                  amplifier
                                       H*                                          LH%
========================  ==========================  ==================  ==================
np:(X^T)^def(X,S&T)/n:X^S  n:X^(Y&(*british(X)))/n:X^Y  n:X^(*amplifier(X))  T@(u/th/(u/th)) \ T
          Z/Z                     (rh/bl)/(rh/bl)              Z/Z          (u/th\(u/th)\(E/bl)
                                 ------------------------------------------>(>B)
                                          n:X^(*amplifier(X)&*british(X))
                                                       rh/bl
          ------------------------------------------------------------->(>B)
          np:(X^T)^def(X,(*amplifier(X)&*british(X))&T)
                                rh/bl
          ------------------------------------------------------------------------------<(<)
                              np:(X^T)^def(X,(*amplifier(X)&*british(X))&T) @ (u/th\(u/th))
                                                   u/th/(u/th)
```

Since the distinction between contrastiveness and newness represents a crucial element of the theory of focus expounded in Chapters 3 and 4, it is also necessary to encode such distinctions in prosodic grammars. For the purposes of generation, the grammar should encode the fact that referentially contrastive items are liable to be more heavily stressed than other items.[11] Moreover, it should handle two contrastive items that occur in the same phrase and require different levels of stress, as discussed in Section 2.3. Example 5.12 illustrates the utterance and phrase level constituents that such a grammar might produce. Recall that particularly emphatic pitch accents are assigned the subscript c. The subscript $[c]$ marks a pitch accent as optionally carrying contrastive weight, while the superscript diacritic "!" marks one or more successively downstepped occurrences of the pitch accent on which it occurs. In order to realize contrastive pitch accents in synthetic speech, theme and rheme categories are assigned relative accent height values.[12] The phrase level and utterance level constituents licensed by this scheme are presented in Example 5.12. The

[11] Here we are appealing to Bolinger's (1989) assertation that accents on so-called contrastive items overshadow other accents in the phrase.

[12] This writer does not believe that contrastive accents are intrinsically different from non-contrastive accents. Furthermore, there is no reliable evidence to suggest that contrastive pitch accents are always realized in a pitch range that is absolutely higher than the pitch range for non-contrastive accents. Bolinger (1989) asserts that the force of contrastive accents is determined by "the degree to which [they] overshadow other accents." For the purposes of generation we simplify this model by setting default levels at which contrastive accents overshadow other accents. Note, however, that when the grammar allows multiple contrastive accents in a phrase, it forces them to be realized at different levels of stress.

corresponding pitch accent and boundary categories are given in Grammar 5.3.

Example 5.12

Utterance Level Constituents:

$$
\begin{array}{llll}
\{\mathrm{L{+}H}^*_{[c]}\}^+\ \mathrm{L(H\%)} & \{\mathrm{H}^*_{[c]}\}^+\ \mathrm{LL\$} & := & u \\
\{\mathrm{L{+}H}^*_{[c]}\}^+\ \mathrm{L(H\%)} & \{\{\mathrm{H}^*_{[c]}\}^+\ \mathrm{LH\%}\}^+ \quad \{\mathrm{H}^*_{[c]}\}^+\ \mathrm{LL\$} & := & u \\
\{\mathrm{H}^*_{[c]}\}^+\ \mathrm{L(L\%)} & \{\mathrm{L{+}H}^*_{[c]}\}^+\ \mathrm{LH\$} & := & u \\
\{\{\mathrm{H}^*_{[c]}\}^+\ \mathrm{LH\%}\}^+ & \{\mathrm{L{+}H}^*_{[c]}\}^+\ \mathrm{LH\$} & := & u
\end{array}
$$

Phrase (Information Structural) Level Constituents:

$$
\begin{array}{ll}
\{\mathrm{H^*}\}^+\ \mathrm{LL\$} & := \quad rh(1) \\
\{\mathrm{H}^*_c\}^!\ \mathrm{LL\$} & := \quad rh(3) \\
\{\mathrm{H^*}\}^+\ \mathrm{L(L\%)} & := \quad u/th(X) \\
\{\mathrm{H}^*_c\}^!\ \mathrm{L(L\%)} & := \quad u/th(X) \\
\{\mathrm{H^*}\}^+\ \mathrm{LH\%} & := \quad rh(3)/rh(X) \\
 & := \quad u/th_2(X)/(u/th_2(X)) \\
 & := \quad u/th(X)\backslash(u/th_2(X)/(u/th_2(X))) \\
\{\mathrm{H}^*_c\}^!\ \mathrm{LH\%} & := \quad rh(3)/rh(X) \\
 & := \quad u/th_2(X)/(u/th_2(X)) \\
 & := \quad u/th(X)\backslash(u/th_2(X)/(u/th_2(X))) \\
\{\mathrm{L{+}H^*}\}^+\ \mathrm{LH\$} & := \quad th(1) \\
\{\mathrm{L{+}H}^*_c\}^!\ \mathrm{LH\$} & := \quad th(2) \\
\{\mathrm{L{+}H^*}\}^+\ \mathrm{L(H\%)} & := \quad u/rh(X) \\
\{\mathrm{L{+}H}^*_c\}^!\ \mathrm{L(H\%)} & := \quad u/rh(X)
\end{array}
$$

Grammar 5.3

Pitch Accent Categories:

H*	:=	$rh(1)/bl$,	$(rh(1)/bl)/(rh(1)/bl)$
H^*_c	:=	$rh(3)/bl$,	$(rh(3)/bl)/(drh(2)/bl)$
$!H^*_c$	:=	$drh(X)/bl$,	$(drh(X)/bl)/(drh(Y)/bl)$, where $Y < X$ and $Y \geq 0$
L+H*	:=	$th(1)/bh$,	$(th(1)/bh)/(th(1)/bh)$
$L+H^*_c$	:=	$th(2)/bh$,	$(th(2)/bh)/(th(2)/bh)$
Ø	:=	X/X,	where X is a variable and
			$\forall Y, Z \quad X \neq Y\backslash Z$

Boundary Categories:

LL$	:=	$T@E \setminus T$
		$E\backslash(E/bl)$
L(L%)	:=	$T@(u/th(Y)) \setminus T$
		$(u/th(Y))\backslash(rh(X)/bl)$
LH%	:=	$T@(rh(X)/rh(Y)) \setminus T$
		$(rh(X)/rh(Y))\backslash(rh(X)/bl)$
	:=	$T@(u/th_2(X)/(u/th_2(Y))) \setminus T$
		$u/th_2(X)/(u/th_2(Y))\backslash(rh(X)/bl)$
	:=	$T@(u/th(X))\backslash(u/th_2(Y)/(u/th_2(Y))) \setminus T$
		$(u/th(X))\backslash(u/th_2(Y)/(u/th_2(Y)))\backslash(E/bl)$
LH$	:=	$T@E \setminus T$
		$E\backslash(E/bh)$
L(H%)	:=	$T@(u/rh(Y)) \setminus T$
		$(u/rh(Y))\backslash(th(X)/bh)$

Note that the category for $\mathbf{H}^*_c$ can only combine with other high pitch accents that are downstepped (i.e. $\mathbf{!H}^*_c$). Likewise, note that downstepped pitch accents can only combine to the right with other downstepped pitch accents. This property of the grammar

encodes the fact that multiple contrastive pitch accents in a phrase are generally realized by different pitch levels (Bolinger 1989).[13]

Having discussed the formal properties of CCGs and their ability to encode the necessary syntactic, semantic and prosodic aspects of language, we now turn to the problem of parsing and generating with CCGs.

5.3 CCG Parsing

Because of CCG's ability to capture the syntactic, semantic, information structural and prosodic aspects of the constituents it licenses, a CCG parser is capable of producing output that is is far richer than simple syntactic trees. That is, a CCG parser can easily construct semantic representations, determine theme/rheme articulations and assign focus within information structural constituents. When a theme has been established in the discourse and there are no salient contrasting themes, the words which realize the theme are often prosodically unmarked. By giving a CCG parser access to the discourse model, prosodically unmarked themes can be identified by matching candidate themes (possible CCG constituents) against the previously established themes. The current implementation works by accepting the match with the most recent established theme in the discourse model. If no such theme exists, the parser takes the longest possible string of prosodically unmarked words as the theme. As new thematic constituents are interpreted by the CCG parser, the discourse model can be updated to include them. In this manner, the CCG parser produces side-effects that may influence the interpretation of subsequent utterances.

The simplest kind of parser for CCGs of the type described in the previous section are bottom-up, shift-reduce parsers.[14] The current implementation employs such a parser that prefers reducing to shifting and makes direct use of the CCG-Prosody theory to derive not only an interpretation of the semantic structure, but also representations for its theme and rheme. While this type of parser is not particularly efficient, its transparency with

[13] This downstepping strategy is effective for realizing multiple contrasts in utterance-final rhemes, but may not apply equally to utterance-initial rhemes.

[14] CCGs are polynomially parsable.

respect to the underlying theory is often useful.

5.4 CCG Generation

Just as the shift-reduce parser sketched above can readily be made to construct the interpretations and information structures shown in the examples, specifically marking themes, rhemes and their foci, so it is relatively easy to do the reverse—to generate prosodically annotated strings from a focus-marked semantic representation of themes and rhemes.

For simplicity, we start by describing the syntactic and semantic aspects of the tactical generator, deferring further discussion of prosody for a moment. Several design options are available, including bottom-up, top-down and semantic head-driven models (Gerdeman and Hinrichs 1990, Shieber and Schabes 1991). We adopt a hybrid approach, employing a basic top-down strategy that takes advantage of the CCG notion of "functional head" to avoid fruitless search.[15] While this technique exhibits some inefficiencies characteristic of a depth–first search, it has several significant advantages. First, it does not rely on a specific semantic representation, and requires only that the semantics be compositional and representable in Prolog. Thus the generating procedure is independent of the particular grammar. This modular character of the system has been very useful in developing the competence grammar proposed in the preceding section, and offers a basis for proving the completeness of the implementation with respect to the competence theory.

The syntactic generation program, also written in Prolog, works as follows. Starting with a syntactic constituent (initially s) and a fully lexicalized semantic representation, we invoke the CCG reduction rules in reverse, as productions, to determine possible subconstituents that can combine to yield the original constituent, invoking the generator recursively to generate the proposed subconstituents.[16]

The base case of the recursion occurs when a category we wish to generate unifies

[15] Hoffman (1995) has developed a bottom-up approach for generating Turkish word orders in Multiset-CCG.

[16] While the backtracking in the generator does allow for multiple lexical realizations for a given input representation, no decisions concerning the appropriateness of competing realizations are made. Such lexical decisions, which are taken to be responsibility of the higher level sentence planner, are considered to be fully encoded in the input. The issue of lexical choice is addressed in more detail in Chapter 6.

with a category in the lexicon. For example, suppose we wish to generate an utterance corresponding to the category $s : walks'(gilbert')$. Since the given category does not unify with any category in the lexicon, the program proposes possible subconstituents by checking the CCG combination rules in some pre-determined order. By the backward function application rule, we might hypothesize that the categories X and $s : walks'(gilbert')\backslash X$ are the subconstituents of $s : walks'(gilbert')$, where X is some variable. If we recursively call the generator on $s : walks'(gilbert')\backslash X$, we find that it unifies with the category $s : walks'(Y)\backslash np : Y$ in the lexicon, corresponding to the lexical item *walks*. This unification forces the complementary category X to unify with $np : gilbert'$, which yields the lexical item *gilbert* when the generator is recursively invoked. Concatenating the results of generating the proposed subconstituents therefore gives the string *"Gilbert walks."*

The top-down nature of the generation scheme has a number of important consequences. First, the order in which we generate the postulated subconstituents determines whether the generation succeeds. Had we chosen to generate X before $s : walks'(gilbert')\backslash X$, we would have entered a potentially infinite recursion, since X unifies with every category in the lexicon. For this reason, our generator always chooses to recursively generate the subconstituent that acts as the functional head before the subconstituent that acts as the argument under the CCG combinatory rules. By strictly observing this principle, we ensure that as much semantic information as possible is deployed, thereby constraining the search space by prohibiting spurious unifications with incorrect items in the lexicon. For this reason, we refer to our generation scheme as a "functional head"-driven, top-down approach.

One disadvantage of the top-down generation technique is its susceptibility to the non-termination problem. If a given path through the search space does not lead to unification with an item in the lexicon, some condition which aborts the path in question at some search depth must be imposed. Note that whenever the CCG function application rules are used to propose possible subconstituents to be recursively generated, the subconstituent acting as the functional head has one more curried argument than its parent. Since we know that in English there is a limit to the number of arguments that a functional category can take, we can abort fruitless search paths by imposing a limit on the number

of curried arguments that a CCG category can possess. The current implementation allows categories with up to three arguments, the minimum needed for constructions involving di-transitive verbs. Note that this strategy does not prohibit the generation of categories whose arguments themselves are complex categories. Thus, we allow categories such as $((s\backslash np)/np)\backslash(((s\backslash np)/np)/np)$ for raised indirect objects, but not categories such as $(((s\backslash np)/np)/np)/np$.

When the CCG composition rule is used to propose possible subconstituents, the subconstituents do *not* have more curried arguments than their parent. Consequently, imposing a bound of the type described above will not necessarily avoid endless recursion in all cases. Suppose, for example that we wish to generate a category of the form s/X, where s is a fully instantiated expression and X is a variable. If the function application rules fail to produce subconstituents that generate the category, we rely on the CCG composition rule to propose the possible subconstituents s/Y and Y/X. Since s/X and s/Y are identical categories except for the renaming of variables, the recursion will continue indefinitely. We rectify this situation by invoking the composition rule only if the original category has an instantiation for both its argument and result. Such a solution imposes limitations on the types of derivations allowed by the system, but retains the simplicity and transparency of the algorithm. Merely imposing a limit on the depth of the recursion provides a more general solution. Examples of the types of sentences that can be generated appear in (74) and (75).

(74)
```
gen(s:(pas^past)^def(x3, (piano(x3)&steinway(x3))&
                         def(x1,(millionaire(x1)&generous(x1))&
                                def(x2,(student(x2)&graduate(x2))&give(x1,x2,x3))))

    ANSWER: the steinway piano was given to the graduate student by the
            generous millionaire.
```

(75)
```
gen(s: (pas^past)^def(x2,piano(x2)&def(x1,(millionaire(x1)&generous(x1))&
                                       give(x1,scott,x2))))

    ANSWER: the piano was given to scott by the generous millionaire.
```

This procedure can immediately be applied to the prosodically augmented grammar.

To do so, we merely enforce the Prosodic Constituent Condition at each step in the generation. That is, whenever a pair of subconstituents are considered (by reversing the CCG combination rules), a pair of prosodic subconstituents are also considered and recursively generated using the prosodic combinatory rules. Examples (76) and (77) illustrate the generation of intonation for the theme and rheme of the utterance "(The OLD patient probably needs)(a LEFT thoracotomy)".[17]

(76) `gen(s:def(x,(patient(x)&*old(x))&probably(needs(x,y)))/np:y @ u/rh).`
`RESULT: the old@lhstar patient probably needs@lhb.`

(77) `gen(np:(x^s)^exists(x,(thoracotomy(x)&*left(x))&s) @ rh).`
`RESULT: a left@hstar thoracotomy@lls.`

Examples (78) and (79) manifest the intonational results of moving the thematic focus among the various propositions in the semantic representation of the theme "The old patient probably needs ...".

(78) `gen(s:def(x,(patient(x)&old(x))& *probably(needs(x,y)))/np:y @ u/rh).`
`RESULT: the old patient probably@lhstar needs@lhb.`

(79) `gen(s:def(x,(*patient(x)&old(x))&probably(needs(x,y)))/np:y @ rh).`
`RESULT: the old patient@lhstar probably needs@lhb.`

5.5 Summary

This chapter presented an intonational competence model in the framework of Combinatory Categorial Grammar (Steedman 1991a). CCG is a mildly context-sensitive grammatical formalism which licenses congruent syntactic, prosodic and information structural constituents, and consequently represents a simplification over competence models of prosody developed in the syntactically more traditional framework of transformational grammar (Chomsky and Halle 1968, Selkirk 1984). Because of the structural similarities among these linguistic entities under CCG, the model for language generation may

[17] The @ symbol separates syntactic categories from their corresponding prosodic categories and lexical items from their pitch/boundary markings.

be considered to involve a single path from high-level semantic representations to spoken utterances, without requiring any additional structures for mapping between surface syntactic forms and intonational domains.

The following issues were examined:

- A formal definition of CCG was provided along with an informal analysis of its potential to generate mildly context-sensitive languages. The combinatory notions of functional application, generalized functional composition and type-raising were introduced and demonstrated by numerous examples.

- The suitability of CCG to describe syntactic phenomena in natural language was exhibited for English constructions. The inclusion of forward composition and type-raising was shown to permit derivations for simple sentences that violate typical right-branching syntactic structure, but generate the same function-argument semantic structure. These non-traditional constituents render CCG useful for handling intonational phrases that often cross traditional syntactic boundaries.

- The interlaced Prosodic and Syntactic CCGs are governed by the Prosodic Constituent Condition, which stipulates that syntactic derivational categories may combine only if their corresponding prosodic categories may also combine.

- A technique for building prosodic CCGs was demonstrated by the three-stage incremental development of such a grammar. The grammars in Section 5.2 encode the necessary level of intonational detail to account for the types of information structure representations developed in Chapter 4. This level of intonational detail in the grammar is instrumental in the development of the natural language generation programs developed in Chapter 6.

- A simple bottom-up shift-reduce parser was presented for deriving semantic and information structural representations from intonationally-annotated sentences.

- A top-down "functional"-head driven algorithm for generating sentences with intonational annotations was introduced. This algorithm, the first such generation

algorithm for CCG, was shown to be applicable to the broader natural language generation task described in Chapter 6.

Because of the congruences among information structure, intonation, and syntax, the CCG formalism and the grammars presented above form an ideal framework for modeling spoken language production. The next chapter adopts this framework for the task of generating contextually-appropriate spoken language from high-level semantic representations.

Chapter 6

Context-Appropriate Intonation for Natural Language Generation

> *It is very difficult to give a collection of synthesized words an overall rhythm and emphasis at the phrase or sentence level, which is important not only to sound good but also to include color, expression, and tone in accordance to both content and intent. Otherwise, the result is a monotonic voice that sounds like a drunken Swede.*
> *Nicholas Negroponte (1995, p. 145)*

In previous chapters, we established the necessity of modeling intonation in context and provided a framework for spoken language production that integrates discourse level information with syntactic, semantic, and information structural constituents. Chapter 4 presented algorithms for building the appropriate information structure representations which in turn can be used by the CCG generation algorithm described in Chapter 5 to produce sentences with context-appropriate prosodic annotations. The present chapter provides computational implementations that demonstrate how the aforementioned algorithms fit together to produce spoken output in a natural language generation framework. While the work presented here does not fully address the range of prosodic problems exhibited by the "drunken Swede," the results illustrate the usefulness of modeling contrast in a meaning-to-speech system.

Two implementations are described in the following sections. First, Section 6.1 presents a program that provides spoken responses to *wh*-questions given as typed input with intonational annotations. This program demonstrates the ability of the model expounded in the previous chapters to generate radically different intonational contours for a single sentence depending on the discourse context (Prevost and Steedman 1993a,1994b). The responses to the questions are generated with respect to a simple knowledge base, the discourse model, and the CCG rules and lexicon.

The second system generates short monologues which describe objects in the knowledge base with respect to the discourse context and varies intonational features appropriately (Prevost 1995). Since this task involves organizing and generating multiple sentences into a coherent discourse segment, it draws on research in both natural language generation and discourse structure. In Section 6.2 several high-level content generation and sentence planning issues are introduced and discussed with respect to various generation frameworks. The chapter concludes with a description of the monologue generation program and its relation to the model of spoken language production set forth in the preceding chapters.

6.1 Implementation I: Query Responses

Although there are two implemented systems described in the present chapter, they serve somewhat different purposes in justifying the appropriateness of the spoken language model presented above. Since the first of these considers only responses to *wh-questions* in narrowly limited contexts, the assignment of themes and rhemes is relatively straightforward. The contexts and questions can also be arranged to highlight contrastive relationships and the associated contrastive theme-foci and rheme-foci in the information structure representations. On the other hand, the monologue task described in Section 6.3 bases decisions concerning information structure and focus on previous utterances it has produced itself rather than direct questions. Consequently, the question/answer task more easily elicits a broad range of possibilities for theme-rheme articulations and focus placement than does the monologue task.

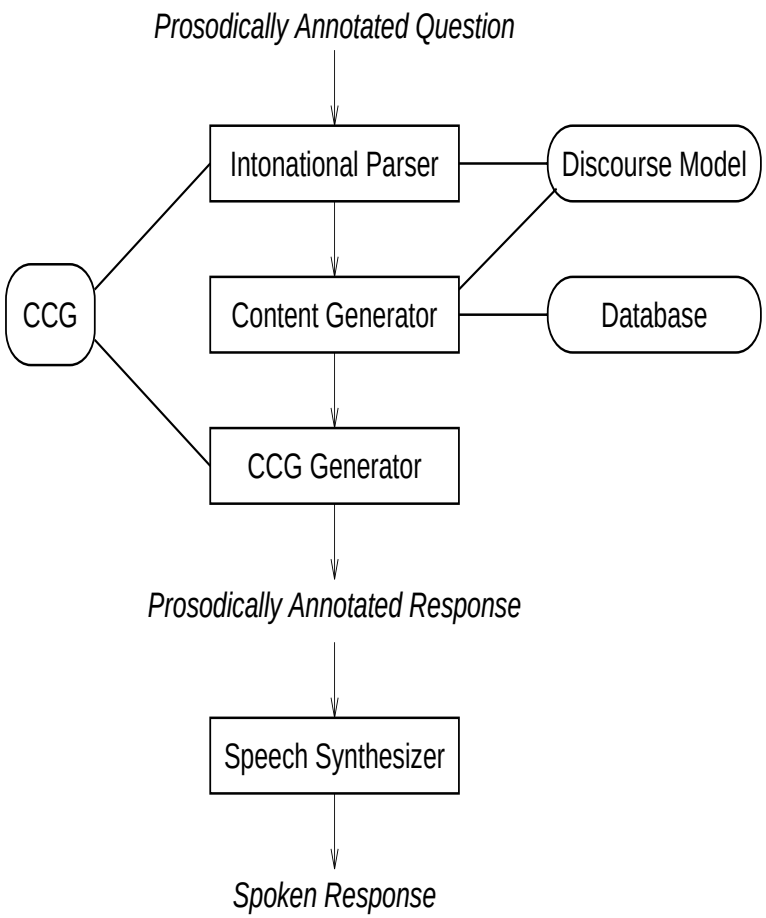


Figure 6.1: An Architecture for Query Response

The architecture of the query response implementation, which is shown in Figure 6.1, identifies the key modules of the system, their relationships to the knowledge base and the underlying CCG grammar, and the dependencies among their inputs and outputs. The process begins with a fully segmented and prosodically annotated representation of a spoken query, as shown in example (80).[1] A simple bottom-up shift-reduce parser, which makes direct use of the combinatory prosody theory described above, identifies the semantics of the question. The inclusion of prosodic categories in the grammar allows the parser to determine the information structure within the question as well, marking theme-focus and rheme-focus items with $\circ$ or $\bullet$ as warranted. The information structure

[1] It should be noted that the program does *not* start with a speech wave, but with a representation that one might obtain from a hypothetical system that translates such a wave into strings of words with Pierrehumbert-style intonation markings.

representations for the question in (80) is illustrated in (81).

(80) I know which components produce MUDDY bass,
but WHICH components produce CLEAN bass?
L+H* LH% H* LL$

(81) Proposition:
$s : \lambda x.component(x)\&produce(x, \bullet clean(bass))$
Theme:
$s : \lambda x.component(x)\&produce(x, \bullet clean(bass))/$
$(s : produce(x, \bullet clean(bass))\backslash np : x)$
Rheme:
$s : produce(x, \bullet clean(bass))\backslash np : x$

In cases where the theme of the question is prosodically unmarked, the parser matches the candidate CCG constituents against themes/rhemes already established in the discourse model. The CCG constituent which unifies with the most recent theme or rheme from the discourse model is taken to be the theme of the question. If no such match exists, the theme of the question is taken to be the longest prosodically unmarked constituent which is syntactically licensed by CCG. In order to facilitate the interpretation of subsequent utterances, the parser updates the discourse model with each newly discovered theme and rheme in the input stream. Consequently, the parser, rather than acting as a purely passive module, produces side effects in the discourse model.

The content generation module, which has the task of determining the semantics and information structure of the response, relies on several simplifying assumptions. Foremost among these is the notion that the rheme of the question is the sole determinant of the theme of the response, including the specification of focus placement. The type of pitch accent (**L+H***) that will eventually mark the theme-focus in the response, however, differs from the pitch accent (**H***) that marks the rheme-focus in the question. The overall semantic structure of the response can be determined by instantiating the variable in the lambda expression corresponding to the *wh*-question with a simple Prolog query. Given the syntactic and focus-marked semantic representation for the response, along with

the syntactic and focus-marked semantic representation for the theme of the response, a representation for the rheme of the response can be worked out from the CCG rules. That is, the rheme of the response is the syntactic and semantic complement of the theme with respect to the overall semantics of the intended utterance. The assignment of rheme-focus in the response (i.e. the variable instantiated by the Prolog query) must be worked out from scratch on the basis of the alternative sets of discourse entities in the knowledge base, as described in the focus assignment algorithm in Chapter 4.

For example, given the question in (80) the content generator produces the following:

(82) Proposition:

$s : produce(\bullet amplifiers, \bullet clean(bass))$

Theme:

$s : produce(x, \bullet clean(bass)) \backslash np : x$

Rheme:

$np : \bullet amplifiers$

Using the output of the content generator, the CCG generation scheme described in Chapter 5 produces a string of words and Pierrehumbert-style markings representing the response, as shown in example (83).

(83) AMPLIFIERS produce CLEAN bass.
H* L L+H* LH$

The final aspect of generation involves translating such a string into a form usable by a suitable speech synthesizer. The current implementation uses the Bell Laboratories TTS system (Liberman and Buchsbaum 1985) as a post-processor to synthesize the speech wave itself, overriding its defaults with the intonational specifications produced by the present program. The TTS system itself is unmodified by the program described above, except for some fine tuning in the lexicon.[2] While TTS is particularly easy to use with Pierrehumbert's notation, the output of the query response program can be easily adapted for other synthesizers.

[2]Previous versions of the present research involved the medical domain of TraumAID, a medical expert system which is under development at the University of Pennsylvania (Webber *et al.* 1992). As expected, due to the specialized domain, several entries had to be added to the TTS pronunciation dictionary.

6.1.1 Results

The query response system described in the previous section produces quite sharp and natural-sounding distinctions of intonation contour in cases such as those in (84)–(91), which manifest the eight basic combinatorial possibilities for pitch accent placement and tune selection produced by the program for the given sentence. These examples illustrate the system's ability to produce appropriately different intonation contours for a single string of words under the control of discourse context. In fact, if the responses in these examples are interchanged, the results sound distinctly unnatural in the given contexts.[3]

Examples (84) and (89) illustrate the necessity of the theme/rheme distinction. Although the pitch accent *locations* in the responses in these examples are identical, occurring on "British" and "treble," the alternation in the theme and rheme tunes is necessary to convey the intended proposition in the given contexts. On the other hand, examples (84) and (87) show that the system makes appropriate distinctions in focus placement within themes and rhemes based on context. Although the responses in these two utterances possess the same intonational tunes, the pitch accent location is crucial for conveying the appropriate contrastive properties.

(84) Q: I know which AMPLIFIER produces clean BASS,

but WHICH amplifier produces clean TREBLE?
L+H* L(H%) H* LL$

A: The BRITISH amplifier produces clean TREBLE.
H* L(L%) L+H* LH$

(85) Q: I know which AMPLIFIER produces MUDDY treble,

but WHICH amplifier produces CLEAN treble?
L+H* L(H%) H* LL$

A: The BRITISH amplifier produces CLEAN treble.
H* L(L%) L+H* LH$

[3]The *waves* files corresponding to the examples in this section are available by anonymous ftp from ftp.cis.upenn.edu, under the directory */pub/prevost/dissertation*.

(86) Q: I know which BRITISH component produces clean BASS,

but WHICH British component produces clean TREBLE?

L+H* L(H%) H* LL$

A: The British AMPLIFIER produces clean TREBLE.

H* L(L%) L+H* LH$

(87) Q: I know which BRITISH component produces MUDDY treble,

but WHICH British component produces CLEAN treble?

L+H* L(H%) H* LL$

A: The British AMPLIFIER produces CLEAN treble.

H* L(L%) L+H* LH$

(88) Q: I know the AMERICAN amplifier produces MUDDY treble,

but WHICH kind of treble does the BRITISH amplifier produce?

L+H* L(H%) H* LL$

A: The BRITISH amplifier produces CLEAN treble.

L+H* L(H%) H* LL$

(89) Q: I know the AMERICAN amplifier produces clean BASS,

but WHICH clean range does the BRITISH amplifier produce?

L+H* L(H%) H* LL$

A: The BRITISH amplifier produces clean TREBLE.

L+H* L(H%) H* LL$

(90) Q: I know the British SPEAKER produces MUDDY treble,

but WHICH kind of treble does the British AMPLIFIER produce?

L+H* L(H%) H* LL$

A: The British AMPLIFIER produces CLEAN treble.

L+H* L(H%) H* LL$

(91) Q: I know the British SPEAKER produces clean BASS,

but WHICH clean range does the British AMPLIFIER produce?

L+H* L(H%) H* LL$

A: The British AMPLIFIER produces clean TREBLE.

L+H* L(H%) H* LL$

6.1.2 Evaluating Results

A formal evaluation of the output shown above has been left to future research because of the complexity of the necessary experimental design. While a simple experiment in which subjects are asked to rate the relative "naturalness" of question-answer pairs would be likely to produce encouraging results, the inclusion of subjective, conscious judgments would make the results difficult to evaluate.[4] Typically such psychological experiments require that subjects not be aware of the variable being tested. The simple experiment mentioned above, however, requires subjects to actively think about the variable under examination, in this case intonation. For this reason, no such experiment has been conducted.

One promising approach to evaluating the output may be an experiment that presents synthetic speech with both contextually-appropriate and contextually-inappropriate intonation, and tests the ability of the subjects to correctly identify referents. Such an experiment would include a comprehension task and a method of manipulating irrelevant variables such as syntax or segmental aspects of the synthesized speech. While such an experiment clearly avoids the problems mentioned above, its complexities, which are beyond the scope of this dissertation, make it better suited for examination by psychologists and psycholinguists.

[4]The assertion that such an experiment would be likely to produce encouraging results is based on informal presentations of the synthesized output at several conferences (Prevost and Steedman 1993a,1993b,1994b,1994a).

6.2 Generation Frameworks

The task of natural language generation (NLG) has often been divided into two stages: content generation, in which high-level goals are satisfied and discourse structure is determined, and surface generation, in which the high-level propositions are converted into sentences. Recently, many NLG researchers have posited the need for an intermediate generation stage, often termed *sentence planning*, in which high-level abstract semantic representations are mapped onto representations that more fully constrain the possible sentential realizations (e.g. Rambow and Korelsky 1992; Reiter and Mellish 1992; Meteer 1991; Elhadad 1992). Given the trend in the NLG community towards the three-tiered approach, this architecture is taken as the model for the present implementation. Since previous research has not included the determination of intonation contours in the three-tiered generation architecture, this issue must be addressed. Recall that we identified two distinct levels of information structure: the theme-rheme articulation and the focus placement within themes and rhemes. In the discourse model described in Chapter 3, coherence is defined as a sequence of relationships between the content of utterance U_{i-1} and the thematic material in utterance Ui. That is, when deciding what to say next, one must consider how an utterance relates to previous utterances and, in particular, what semantic information is shared between consecutive utterances. Since the process of organizing a discourse at the highest level must be concerned with the overall discourse coherence, the division of high-level propositions into theme and rheme must be handled at the content determination phase of the three stage NLG model described above. In this manner, the content generator ensures that consecutive pairs of utterances share semantic material whenever such an ordering of utterances exists.

The process of determining foci within themes and rhemes can be divided into two tasks: determining which discourse *entities* or *propositions* are in focus, and determining how their linguistic realizations should be marked to convey that focus. The first of these tasks can be handled in the content phase of the NLG model described above. The second of these tasks, however, relies on information, such as the construction of referring expressions, that is often considered the domain of the sentence planning stage. For example, although two discourse entities e_1 and e_2 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. Recent generation research employs a level of representation for referring expressions that is separate from the intended semantic content and consequently is determined in the sentence processing stage (Dale and Haddock 1991). Similarly, the present approach resolves issues of contrastive focus assignment at the sentence processing stage.

During the content generation phase, the content of the utterance is planned based on the previous discourse. There are two general bases for constructing the plans: *schemata* and *rhetorical structure*. The *schemata* approach, put forth by McKeown (1985), involves selecting a proper template for conveying the desired message and fleshing out the details based on the available information. Because the templates are essentially written as regular expressions with optional and potentially repeatable parts, there is a great deal of flexibility and the resulting texts do not sound repetitive. Unfortunately, a given set of templates may not apply equally well across multifarious domains. For example, a set of templates designed for a medical application may not produce the most natural texts for a business application.

Although the template-based system of McKeown is generally the most useful for real applications, other researchers have seen the need to incorporate more flexibility into the generation process. By employing rhetorical structure theory (RST, Mann and Thompson 1986), Hovy (1988,1993) attempts to organize texts by identifying rhetorical relations between clause-level propositions from a knowledge base. While this approach is theoretically motivated, it exhibits a number of problems in practice, such as the difficulty in defining a finite set of rhetorical predicates.

Another novel approach to the high-level generation planning problem is offered by Sibun (1991,1992). In her approach, plans for describing architectural layouts are generated without producing discourse trees. Propositions are linked to one another not by rhetorical relations or pre-planned templates, but rather by physical and spatial properties represented in the knowledge-base.

The present framework for organizing the content of a monologue is a hybrid of all three

approaches mentioned above. The implementation, which is presented in the following section, produces descriptions of objects from a knowledge base with context-appropriate intonation that makes proper distinctions of contrast between alternative, salient discourse entities. Certain constraints, such as the requirement that objects be identified or defined at the beginning of a description, are reminiscent of McKeown's schemata. Rather than imposing strict rules on the order in which information is presented, the order is determined by domain specific knowledge (as in Sibun 1991, 1992), the communicative intentions of the speaker, and beliefs about the hearer's knowledge. Finally, the system includes a set of rhetorical constraints that may rearrange the order of presentation for information in order to make certain rhetorical relationships salient. While this approach has proven effective in the present implementation, further research is required to determine its usefulness for a broader range of discourse types.

6.3 Implementation II: Monologue Generation

While the query response system described in Section 6.1 demonstrates the ability of the theory laid out in the previous chapters to model intonation based on context, its reliance on the explicit marking of the intonational contour in the question greatly diminishes the practical applications of the system. Moreover, the query system models context in its most limited sense—the relationship between questions and answers—and consequently does not demonstrate the underlying theory's ability to handle the more subtle types of contrastive situations that arise in real discourse. The monologue generation program described here attempts to address these issues by producing contextually-appropriate intonation contours based on several connected sentences concerning a single object. The system exhibits the ability to intonationally contrast alternative entities and properties that have been explicitly evoked in the discourse even when they occur with several intervening sentences.

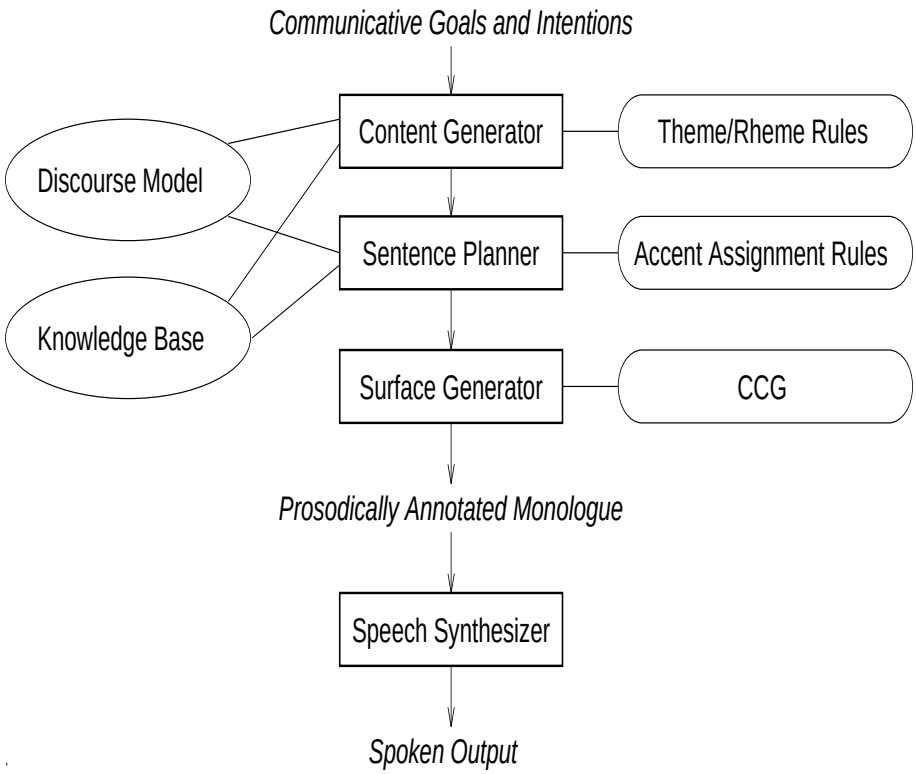


Figure 6.2: An Architecture for Monologue Generation

6.3.1 Content Generation

The architecture for the monologue generation program is shown in Figure 6.2, in which arrows represent the computational flow and lines represent dependencies among modules. The remainder of this section contains a description of the computational path through the system with respect to a single example. The input to the program is a goal to describe an object from the knowledge base, which in this case contains a variety of facts about hypothetical stereo components.[5] In addition, the input provides a communicative intention for the goal which may affect its ultimate realization, as shown in (92). For example, given the goal `describe(x)`, the intention `persuade-to-buy(hearer,x)` may result in a radically different monologue than the intention `persuade-to-sell(hearer,x)`.

(92) GOAL: `describe e1`

INPUT: `generate(intention(believe(h1,good-to-buy(e1)))`

[5]The domain of the knowledge base is completely divorced from the generation algorithm and can be easily modified to encompass other similar domains.

Information from the knowledge base is selected to be included in the output by a set of relations (assumed to be mutually believed by the speaker and the hearer) that determines the degree to which knowledge base facts and rules support the communicative intention of the speaker. Consequently, if the program's intention is to persuade the hearer to buy a particular object, it will offer those propositions that it believes the hearer will consider most important in making a purchasing decision and eschew less relevant propositions. The result is that the output of the system may vary for different hearers. In addition, the system models the relative degree to which certain facts in the knowledge base support the intention of the speakers. For example, suppose the system "believes" that conveying the proposition in (93) moderately supports the intention of making hearer `h1` want to buy `e1`, and further that the rule in (94) is known by `h1`.[6]

(93)
```
believe(h1, holds(rating(X, powerful)))
```

(94)
```
holds(rating(X, powerful), T) :-
        holds(produce(X, Y), T),
        holds(isa(Y, watts-per-channel), T),
        holds(amount(Y, Z), T),
        number(Z),
        Z >= 100.
```

The program then consults the facts in the knowledge base, verifies that the property does indeed hold and consequently includes the corresponding facts in the set of properties to be conveyed to the hearer, as shown in (95).

(95)
```
holds(produce(e1,  e7)).
holds(isa(e7, watts-per-channel)).
holds(amount(e7, 100)).
```

As described in Section 6.2, the content generation module organizes the relevant information from the knowledge base using a hybrid of three text planning paradigms: schemata (McKeown 1985), rhetorical structure theory approaches (RST, Hovy 1988,1993),

[6]The rule is shown in the standard Prolog Horn clause representation. The variable `T` encodes a rather simplistic notion of tense in the knowledge base.

and a domain structure approach (Sibun 1991,1992).[7] The content generator starts with a simple description template that specifies that an object is to be explicitly identified or defined before other propositions concerning it are put forth. Other relevant propositions concerning the object in question are then linearly organized according to beliefs about how well they contribute to the overall intention (hence the analogy to the domain structured generation of Sibun 1992). Finally, a small set of rhetorical predicates rearranges the linear ordering of propositions so that sets of sentences that stand in some interesting rhetorical relationship to one another will be realized together in the output. These rhetorical predicates assist in maintaining the coherence of the output. For example, the *conjunction* predicate specifies that propositions sharing the same theme or rheme be realized together in order to avoid excessive topic shifting. The *contrast* predicate specifies that pairs of propositions that explicitly contrast with one another be realized together. Other rhetorical constraints (e.g. temporal orderings) can easily be added to the system, but have not been necessary for the current domain. The result is a set of properties roughly ordered by the degree to which they support the given intention.

Since each property in the ordered set may contain several discourse entities among its arguments, the content generator also needs to ensure that it conveys enough information about all of the discourse entities so that reference will succeed for the hearer. This is accomplished by invoking the content generator recursively for each item `x` with the goal `describe x` and the simple intention of informing the hearer of what `x` refers to. After completing this step, the information to be conveyed by the program is organized as a hierarchical relationship among discourse entities, as shown in Example 6.1. In this example, the top-level propositions were selected by the program because the hearer (`h1`) is believed to be interested in the design of the amplifier and the reviews the amplifier has received. Moreover, the belief that the hearer is interested in buying an expensive, powerful amplifier justifies including information about its cost and power rating. Different

[7] The schemata approach provides a template for the type of monologue being generated, in this case a "description." While this type of approach is relatively straightforward to implement, a given template may not be generalizable over different domains. The RST approach attempts to organize text by identifying rhetorical relations between clause-level propositions from the knowledge base. While this approach maximizes flexibility, it also may dramatically increase search space. Furthermore, identifying a set of rhetorical relations that covers a wide range of examples is by no means a trivial task.

sets of propositions would be generated for other (perhaps thriftier) hearers. Additionally, note that the propositions `praise(e4,e1)` and `revile(e5,e1)` are combined into the larger proposition `but(praise(e4,e1), revile(e5,e1))`. This is accomplished by the rhetorical constraints that determine the two propositions to be contrastive because `e4` and `e5` belong to the same set of alternative *entities* in the knowledge base and `praise` and `revile` belong to the same set of alternative *propositions* in the knowledge base.

Example 6.1

```
de-tree(
   props(e1,
      [holds(defn(isa(e1,amplifier))),
       holds(design(e1,solid-state),pres),
       holds(cost(e1,e9),pres),
       holds(produce(e1,e7),pres),
       holds(but(praise(e4,e1),revile(e5,e1)),past)]),
   supports(e1,
      [de-tree(
         props(e9,
            [holds(isa(e9,dollar-amount),pres),
             holds(amount(e9,800),pres)]),
         supports(e9,[])),
       de-tree(
         props(e7,
            [holds(isa(e7,watts-per-channel),pres),
             holds(amount(e7,100),pres)]),
          supports(e7,[])),
       de-tree(
         props(e4,
            [holds(isa(e4,journal),pres),
             holds(subject(e4,stereo),pres)]),
         supports(e4,[])),
       de-tree(
         props(e5,
            [holds(isa(e5,journal),pres),
             holds(subject(e5,stereo),pres)]),
         supports(e5,[]))]))
```

The next phase of content generation recognizes the dependency relationships between the properties to be conveyed based on shared discourse entities. This phase, which represents an extension of the rhetorical constraints, arranges propositions to ensure that

consecutive utterances share semantic material (i.e. the theme of the second utterance in a sequence). This rule, which in effect imposes a strong bias for centering's *continue* and *retain* transitions, as described in Section 3.6, determines the theme-rheme segmentation for each proposition.[8]

The careful arrangement of the propositions to ensure the sharing of semantic material between consecutive utterances is similar to the approach used by McKeown, Kukich and Shaw (1994). Their PLANDoc system, which produces summaries of the activities of telephone route planning engineers, arranges messages that share information together in the output. In this manner, conjunction and ellipsis can be used to efficiently convey the information and avoid unwarranted repetition.

6.3.2 Sentence Planning

After the coherence constraints from the previous section are applied, the sentence planner is responsible for making decisions concerning the form in which propositions are realized. This is accomplished by a simple set of rules, which includes the following:

(96) a. Definitional `isa` properties are realized by the matrix verb.
b. Other `isa` properties are realized by nouns or noun phrases.
c. Top-level properties (the non-embedded properties in Example 6.1) are realized by the matrix verb.
d. Embedded properties are realized by adjectival modifiers if possible and otherwise by relative clauses.

Example 6.2 illustrates part of the result of applying these rules to the output of the content generator.

[8]Centering *shifts* are rarely necessitated for such short monologues which describe a single object. Indeed, shifts are only generated when the propositions concerning other entities mentioned in the monologue cannot fit into a single sentence. Since the recursive invocation of the algorithm for describing other entities only produces enough information for reference to succeed, this situation seldom occurs.

Example 6.2

```
[node(e1,
   nouns([]),
   adjs([]),
   clauses(
      [infostruc(
         sem(defn(isa(e1,c1))),
         th(e1),
         rh(x^isa(x,c1)))]),
   sub(
      [[node(c1,
           nouns([isa(c1,amplifier)]),
           adjs([design(c1,solid_state)]),
           clauses([]),sub([]))]])),
 node(e1,
   nouns([]),
   adjs([]),
   clauses(
      [infostruc(
         sem(cost(e1,e9)),
         th(e1), rh(x^cost(x,e9))),
       infostruc(
         sem(and(produce(e1,e7))),
         th(e1),
         rh(x^and(produce(x,e7))))]),
   sub(
      [[node(e9,
          nouns([isa(e9,dollar_amount)]),
          adjs([amount(e9,compound(eight,hundred))]),
          clauses([]),sub([]))],
       [node(e7,
          nouns([isa(e7,watts_per_channel)]),
          adjs([amount(e7,compound(one,hundred))]),
          clauses([]),sub([]))]])), . . . ]
```

Although the rules in (96) are sufficient for producing text that exercises a broad range of intonational possibilities, other generation approaches have incorporated more detailed rules that might reasonably replace (96)c and (96)d. Robin (1993, 1994), and Robin and McKeown (1993) employ revision-based techniques that allow additional information to be added at the sentence planning level. Such *Additional Deep Syntactic Specifications* (ADSSs) are similar to the embedded properties in Example 6.2. Robin's approach allows these embedded properties to be hooked into an existing Deep Syntactic Specification (DSS) in various locations and in a variety of forms. For example, the DSS corresponding to the sentence "John Stockton scored 27 points" might be elaborated to "John Stockton scored *a career high* 27 points."

While there are certainly a number of linguistically interesting aspects to the sentence planner, the most important aspect for the present purposes is the determination of theme-foci and rheme-foci. The focus assignment algorithm employed by the sentence planner, which has access to both the discourse model and the knowledge base, works exactly as set forth in Chapter 4. First, each property or discourse entity in the semantic and information structural representations is marked as either previously mentioned or new to the discourse. This assignment is made with respect to both the discourse entity list (DEList) described in Section 4.2 and a similar structure for evoked properties. Certain aspects of the semantic form are considered unaccentable because they correspond to the interpretations of closed-class items such as function words. Items that are assigned focus based on this implementation of the previous mention heuristic are assigned the $\circ$ focus operator, as shown in (97).

(97)

Semantics:	$defn(isa(\circ e1, \circ c1))$
Theme:	$\circ e1$
Rheme:	$\lambda x.isa(x, \circ c1)$
Supporting Props:	$isa(c1, \circ amplifier)$
	$\circ design(c1, \circ solid-state)$

The second step in the focus assignment algorithm checks for the presence of contrasting propositions in the ISStore, a structure that stores a history of information structure representations. Propositions are considered contrastive if they contain two contrasting

pairs of discourse entities, or if they contain one contrasting pair of discourse entities as well as contrasting functors. Both of these situations are illustrated in (98) below.

(98) a. John likes Mozart, but Mary likes Beethoven.

b. John likes Mozart, but he hates Beethoven.

Discourse entities are determined to be constrastive if they belong to the same set of alternatives in the knowledge base, where such sets are inferred from the `isa`-links that define class hierarchies. As more and more `isa`-links are traced in search of a common class, the size of the set of alternatives increases. Given the assertion put forth by Bolinger (1961) and Schmerling (1976) that the contrastive effect is inversely proportional to the size of the set of alternatives, the effect is similarly related to the distance (measured in `isa`-links) of the entities in the knowledge base. While the present implementation only considers entities with the same parent or grandparent class to be alternatives for the purposes of contrastive stress, a graduated approach that entails degrees of contrastiveness may also be possible.

The effects of the contrastive focus algorithm are easily shown by examining the generation of an utterance that contrasts with the utterance shown in (97). That is, suppose the generation program has finished generating the output corresponding to the examples in (92) through (97) and is assigned the new goal of describing entity `e2`, a different amplifier. After applying the second step on the focus assignment algorithm, contrasting discourse entities are marked with the $\bullet$ operator, as shown in (99). Since `e1` and `e2` are both an instances of the class `amplifiers` and `c1` and `c2` both describe the class `amplifiers` itself, these two pairs of discourse entities are considered to stand in contrastive relationships.

(99)

Semantics:	$defn(isa(\bullet e2, \bullet c2))$
Theme:	$\bullet e2$
Rheme:	$\lambda x.isa(x, \bullet c2)$
Supporting Props:	$class(c2, amplifier)$
	$design(c2, \circ tube)$

While the second step of the algorithm determines which abstract discourse entities and properties stand in contrast, the third step uses the contrastive focus algorithm

described in Section 4.2 to determine which elements need to be contrastively focused for reference to succeed. For example, although the representation in (99) specifies that `e2` stands in contrast to some other entity, it is the property of `e2` having a tube design rather than a solid-state design that needs to be conveyed to the hearer. After applying the third step of the focus assignment to (99), the result appears as shown in (100), with '"tube" contrastively focused as desired.

(100)

Semantics:	$defn(isa(\bullet e2, \bullet c2))$
Theme:	$\bullet e2$
Rheme:	$\lambda x.isa(x, \bullet c2)$
Supporting Props:	$isa(c2, amplifier)$
	$design(c2, \bullet tube)$

The final step in the sentence planning phase of generation is to compute a representation that can serve as input to the CCG generator described in Section 5.4. This module contains a straightforward procedure for building referring expressions that follows the guidelines stated in (101).

(101) a. If entity e is part of the theme of the current utterance, is realized in the previous utterance and has no restrictive modifiers, then realize e as a pronoun.

b. If entity e has an explicit name that is known to the hearer, realize e by its name. If e also has modifiers that must be realized, include them in a parenthetical clause.

c. Otherwise, determine the definiteness of e and realize it as a noun phrase with appropriate modifiers.

The resulting semantics for the focused form shown in (100) is presented below in (102).

(102)

Theme:	$np(3,s) : (e1 \hat{} S) \hat{} def(e1, \bullet x5(e1) \& S)@u/rh(1)$
Rheme:	$s : (act \hat{} pres) \hat{} indef(c1, (amplifier(c1) \& \bullet tube(c1))$
	$\& isa(e1, c1)) \backslash np(3,s) : e1@rh(1)$

Given the focus of this dissertation, the sentence planner is concerned mainly with determining the information structural correlates (theme- and rheme-foci) of intonational

patterns of accentuation. A number of other issues that are generally considered the responsibility of the sentence planner are given only minimal treatment here. For example, the issue of lexical choice for verbs is handled by directly mapping high-level propositions onto verbs in the lexicon. For noun phrases, lexical choices are made with respect to the algorithm in (101) which constructs appropriate referring expressions. Other researchers (Reiter 1991; Smadja and McKeown 1991; Elhadad and Robin 1992; Elhadad, McKeown and Robin 1996) have studied the problem of lexical choice in generation frameworks with encouraging results. For example, the STREAK system (Robin 1994) is able to produce the sentences in (103)a and (103)b, depending on whether the *manner* constraint is to be conveyed by the prepositional phrase or the verb. Such techniques do not conflict with the information structural aspects of the present sentence planner, and therefore stand to be incorporated during future research.

(103) a. The Denver Nuggets beat the Boston Celtics with a narrow margin, 102–101.
b. The Denver Nuggets edged out the Boston Celtics 102–101.

6.3.3 Surface Generation and Results

Given the output of the sentence planner, the surface generation module described in Section 5.4 consults a CCG grammar which dictates the generation of both the syntactic and prosodic constituents, and produces a string of words and the appropriate prosodic annotations, as shown in (104). The output of this module is easily translated into a form suitable for a speech synthesizer, which produces spoken output with the desired intonation.[9]

(104) The X5 is a TUBE amplifier.
$L+H^*_c$ L(H%) H^*_c LL\$

The modules described above and shown in Figure 6.2 are implemented in Quintus Prolog. The system produces the types of output shown in (105) and (106), which should be interpreted as a single (two paragraph) monologue satisfying a goal to describe two

[9]The system currently uses the AT&T Bell Laboratories TTS system, but the implementation is easily adaptable to other synthesizers.

different objects.[10] Note that both paragraphs include very similar types of information, but radically different intonational contours, due to the discourse context. In fact, if the intonational patterns of the two examples are interchanged, the resulting speech sounds highly unnatural. Moreover, the examples show that contrastive focus can occur on items that are well established in the discourse, as exemplified by the contrastive accents placed on the pronoun "it" in (106). The complete and somewhat verbose output of the program for examples (105) and (106) can be found in Appendix B.[11]

(105) a. Describe the x4.

b. The X4 is a SOLID-state AMPLIFIER.
L+H* L(H%) H* H* LL$

It COSTS EIGHT HUNDRED DOLLARS,
H* H* H* H* LL%

and PRODUCES ONE hundred watts-per-CHANNEL.
H* H* H* LL$

It was PRAISED by STEREOFOOL, an AUDIO JOURNAL,
H^*_c $!H^*_c$ LH% H* H* LH%

but was REVILED by AUDIOFAD, ANOTHER audio journal.
H^*_c $!H^*_c$ LH% H* LL$

[10] Recall that the resulting intonation will assign slightly higher pitch to accents bearing the subscript c (e.g. H^*_c), which mark contrastive focus as determined by the algorithm in Section 4.

[11] The *waves* files corresponding to these examples can be acquired by anonymous ftp from ftp.cis.upenn.edu, under the directory */pub/prevost/dissertation*.

(106) a. Describe the x5.

b. The X5 is a TUBE amplifier.

L+H^*_c L(H%) H^*_c LL$

IT costs NINE hundred dollars,

L+H^*_c L(H%) H^*_c LH%

produces TWO hundred watts-per-channel,

H^*_c LH%

and was praised by Stereofool AND Audiofad.

H^*_c LL$

Several aspects of the output shown above are worth noting. First, the decision concerning whether or not to include certain information (propositions) in the output is based on how well that information supports the given communicative intention. Recall that the intention for both of the descriptions shown in (105) and (106) is to convince the hearer to purchase the audio components. The program's model of which information is considered by the hearer to be important for making a positive purchasing decision determines the propositions that are put forth in the utterances. Initially, the program assumes that the hearer has no specific knowledge of any particular objects in the knowledge base. Note however, that every proposition put forth by the generator is assumed to be incorporated into the hearer's set of beliefs. Consequently, the descriptive phrase "an audio journal" which is new information in the first paragraph is omitted from the second. Additionally, when presenting the proposition '*Audiofad is an audio journal*', the generator is able to recognize the similarity with the corresponding proposition about Stereofool (i.e. both propositions are abstractions over the single variable open proposition '*X is an audio journal*). The program therefore interjects the *other* property and produces "another audio journal."

Several aspects of the contrastive intonational effects in these examples also deserve attention. Because of the content generator's use of the rhetorical contrast predicate, items are eligible to receive stress in order to convey contrast before the contrasting items are even mentioned. This phenomenon is clearly illustrated by the clause "PRAISED by

STEREOFOOL" in (105), which is contrastively stressed before "REVILED by AUDIO-FAD" is uttered. Such situations are produced only when the contrasting propositions are gathered by the content planner in a single invocation of the generator and identified as contrastive when the rhetorical predicates are applied.

The contrastive stress on "and" in (106) is due to the application of the conjunction rhetorical predicate. First, during the content generation phase, the two propositions `praise(stereofool, x5)` and `praise(audiofad, x5)` are merged into the single proposition `praise(conj(audiofad, stereofool), x5)`. That is, the new entity `conj(audiofad, stereofool)` is created and inserted into the discourse model. The initial set of alternatives for the focus placement algorithm described in Chapter 4 includes the new entity as well as its components. The initial list of properties includes *stereofool*, *audiofad* and *conj*. Although restricting the *RSet* based on either the *stereofool* or the *audiofad* property would eliminate a single element of the *RSet*, restricting on the *conj* property is required in order to eliminate all but the conjoined alternative from the set. Since the *conj* property alone can eliminate the other alternatives, it alone receives contrastive focus. Hence only the word "and" is stressed in the final output.

6.3.4 The Range of Examples

The implementation described above works for a range of examples that fall under the category of descriptions. Additionally, the system displays a large degree of flexibility in each of the following areas:

(107) a. Selection of information for presentation is based on the program's model of the hearer's beliefs and the relationship between those beliefs and the program's communicative intention.

b. Rhetorical constraints allow for information to be packaged based on inter-clause and inter-utterance semantic relationships (e.g. contrast).

c. Intonational phrasing and accentual decisions are made based on the cumulative discourse context exactly as set forth in Chapter 4.

d. Given (i.e. previously mentioned) information is eligible to receive stress when it stands in contrast to some other previously mentioned information. This is particularly striking in examples where pronouns receive stress. No other intonation generator produces such effects.

The examples in (108) through (110) illustrate the first point listed above. Each of these examples is a discourse-initial monologue about the same entity in the knowledge base, but is geared toward a different hearer. Since hearer `h1` in example (108) is believed to be interested in the distinctive characteristics of the given amplifier rather than its functionality or design, he is informed of its unique history. Hearer `h2` in example (109), on the other hand, is believed to be more concerned with the functionality and is consequently informed of the appropriate details. Note that hearer `h2` is informed of Alexander Graham Bell's occupation because she is not "believed" by the program to know this information. Finally, hearer `h3` in (110) is given very little information because he is believed to be interested only in modern high-end stereo systems.

(108) the X1 is an ANTIQUE AMPLIFIER.
L+H* L(H%) H* H* LL$
it was designed by Albert EDISON,
H^*_c $!H^*_c$ L(H%)
but was BUILT by Alexander Graham BELL.
H^*_c $!H^*_c$ LL$
it COSTS TWO HUNDRED DOLLARS.
H* H* H* H* LL$

(109) the X1 is a TUBE AMPLIFIER.
L+H* L(H%) H* LL$
it COSTS TWO HUNDRED DOLLARS,
H* H* H* H* L(H%)
and PRODUCES FIVE watts per CHANNEL.
H* H* H* LL$
it was BUILT by Alexander Graham BELL, an INVENTOR.
H* H* L(H%) H* LL$

(110) the X1 is a TUBE AMPLIFIER.
L+H* L(H%) H* LL$

The examples below, which are to be taken as consecutive monologues, illustrate the ability of the program to contrast various objects from the knowledge base. Note that since the program is describing two different objects, slightly different types of information are conveyed even though both monologues are intended for the same hearer. Only those propositions that stand in direct contrast to one another are realized with particularly emphatic accents.

(111) the C5 is a BRITISH CD player.
L+H* L(H%) H* H* LL$
it COSTS NINE HUNDRED DOLLARS,
H* H* H* H* L(H%)
and PLAYS FIVE CDs,
H* H* H* LL$

(112) the X4 is a AMERICAN AMPLIFIER.
L+H^*_c L(H%) H^*_c !H^*_c LL$
IT costs EIGHT hundred dollars,
L+H^*_c L(H%) H^*_c L(H%)
and PRODUCES ONE hundred watts per CHANNEL.
H* H* H* LL$

6.4 Limitations

Although the results presented above support the theoretical aspects of the model of spoken language generation espoused in the previous chapters, the computational implementation of the model is limited in a number of ways. Most importantly, the implementation places restrictions on what constitutes a set of alternative entities from the knowledge base. As it stands, the program only considers entities with the same parent or grandparent class to be considered alternatives. As a result, the program is unable to produce output of the type shown in (113), which requires semantically disparate items to be contrasted, unless the knowledge base is specifically engineered to handle such examples.

(113) Mary took John to a hockey game for his birthday, but he didn't seem very pleased.
While HE intently watched the CLOCK, SHE watched the GAME.

Since "game" is mentioned in the first sentence of (113), the accent on "game" on the second sentence must be computed on the basis of alternative sets rather than the previous mention heuristic. In order for the present algorithm to include "clock" and "game" in the contrast set, the knowledge base must include some class which fairly directly links the two objects (e.g. "things in a sports arena"). Consequently, the knowledge representation must be extremely robust in order to account for such occurrences of contrastive accentuation.

As described above, the generator also makes strict limitations on lexical choice, mapping high level concepts directly onto words whenever possible. Note however, that the model presented here is completely compatible with other more sophisticated lexical choice

algorithms (cf. Smadja and McKeown 1991; Elhadad and Robin 1992; Elhadad, McKeown and Robin 1996). In future research, the focus placement algorithm can be easily adapted to operate on the types of intermediate lexicalized structures produced by these generators.

The degree to which the theory and implementation rely on the selection and size of the domain is a somewhat more complex issue. Different domains may require different sets of rhetorical predicates for organizing information in the content planning phase of generation. For example, narrative domains may require temporal constraints to determine the appropriate ordering for events. As new rhetorical constraints are added, the processing time necessarily increases. Since each constraint requires each proposition (from the set selected by the content planner for conveyance) to be compared to every other proposition in the set, the total number of comparisons for each constraint is polynomial in the size of the set. Moreover, before a proposition is even selected to be included in the text, it must be compared to the set of facts which describe the program's "beliefs" about what the hearer considers important (with respect to the underlying intention of the speech act). Consequently, as the size of the knowledge base grows, the number of comparisons increases polynomially in the size of the knowledge base.

Although the size of the knowledge base adversely affects runtime and space complexity (albeit only polynomially), increasing the size of the associated lexicon has very little consequence. Since the input to the surface generator is assumed to be a fully lexicalized logical form, the program first performs a linear scan of the lexicon, collecting only those items that are necessary for the algorithm to halt with a solution. Therefore there can be no spurious unifications with incorrect items in the lexicon. Moreover, the runtime of the surface generator is unaffected except for the initial linear scan. Given these modest limitations, the theory and implementation described above are particularly applicable to applications involving tightly constrained domains, such as interfaces to expert systems.

6.5 Summary

This chapter presented two implementations that demonstrate the usefulness of the theory expounded in the previous chapters for modeling spoken language production. The first implementation provides spoken responses to *wh*-questions given as typed input with intonational annotations. This program has the ability to generate radically different intonational contours for a single sentence depending on the discourse context. The responses to the questions are generated with respect to a simple knowledge base, the discourse model, and the CCG rules and lexicon.

The second implementation generates paragraph-length, spoken monologues concerning objects in a simple knowledge base. The process of natural language generation, in accordance with much of the recent literature in the field, is divided into three processes: high-level content planning, sentence planning, and surface generation. Two points concerning the role of intonation in the generation process are emphasized. First, since intonational phrasing is dependent on the division of utterances into theme and rheme, and since this division relates consecutive sentences to one another, matters of information structure (and hence intonational phrasing) must be largely resolved during the high-level planning phase. Second, since accentual decisions are made with respect to the particular linguistic realizations of discourse properties and entities (e.g. the choice of referring expressions), these matters cannot be fully resolved until the sentence planning phase.

The primary purpose of the implementation is to test the hypothesis that the representations involved in language production, from abstract semantic propositions to surface phonological and syntactic forms, can be treated as structurally congruent. As a consequence of this hypothesis, spoken language generation can be viewed as a straight-line process, as exemplified by the architecture shown in Figure 6.2. The secondary purpose of the implementation is to provide a platform for incrementally testing and refining the competence model of intonation embodied by the CCG categories and rules.

Chapter 7

Applications and Related Research

The principal tenet underlying this dissertation is that intonation and semantics are inextricably linked. A number of researchers have begun to explore the semantics of other forms of extra-linguistic communication as well. Facial expressions, in particular, convey beliefs, intentions, attitudes and finer semantic nuances similar to some of those conveyed by intonation. Therefore, we examine how facial expressions and intonational features act together to communicate a message in a specific context.

This chapter contains a brief summary of some of the related work on the meaning of facial expressions and its relations to the intonational theory and computational model described in the previous chapters. Section 7.4 provides a computational implementation that links the monologue generator described in Chapter 6 to the task of automatically producing contextually-appropriate facial animations.

It should be noted that some of the theories of facial expression discussed below have been contested. The purpose of the implementation described in this chapter, however, is not to advocate any particular theory or assume any theory as a plausible model for describing human facial expression. Rather, the intent of the present implementation is to provide a computational platform for refining and evaluating some of the hypotheses discussed below.[1]

[1]The research and implementation described in this chapter represent a joint effort with Catherine

7.1 Facial Expressions

Thoughts are conveyed with both words and facial expressions. That is, facial actions, such as smiling, raising the eyebrows, and wrinkling the nose, often co-occur with a verbal message. Some facial expressions accompany the flow of speech and are synchronized at the verbal level, punctuating accented phonemic segments and pauses. Other facial expressions may substitute for a word or string of words, or emphasize the spoken message. Facial expressions do not necessarily occur randomly, and have been hypothesized to be synchronized to one's own speech, or to the speech of others (Condon and Ogston 1971, Kendon 1974).

During discourse, the behavior of a person and the presence or absence of feedback by her affect the behavior of the other conversational participants. A conversation consists of the exchange of meaningful utterances and associated physical manifestations. For example, the speaker may punctuate or reinforce her speech by nodding the head, smiling or using a variety of other facial gestures. Likewise, the listener can interact by smiling, vocalizing or shifting his gaze to participate in the conversation. Facial expressions and gaze behavior are important aspects of the interaction because they serve to control the flow of speech and regulate the exchange of speaking-turn. Such expressions depend heavily on the relationship between the discourse participants, as well as their individual personalities, emotions, attitudes and social identities. All movements, both conscious and unconscious, influence the other participants in a conversation.

7.1.1 Synchronism

One hypothesis concerning the link between intonational properties and facial expressions (and also body movements) is that they tend to occur in synchrony (Bull and Connelly 1985, Condon and Ogston 1971, Condon and Sander 1974, Gatewood and Rosenwein 1981). "Synchrony" implies that changes occurring in speech and in body movements appear at the same time. Thus, there should be synchrony at the boundaries (i.e., at the end of an intonational phrase or when a gesture occurs) establishing a congruence between the patterns of speech and gesture. As described by Condon and Ogston (1971), "the body

Pelachaud. A preliminary version of the present text has appeared in (Pelachaud and Prevost 1995b).

parts change direction of movement and sustain direction of movement, and the boundaries formed by the clusterings of such changing-and-sustaining-together, are isomorphic with the articulatory (as against abstracted segments) transformations of speech." For example, as a word begins to be articulated, an eye blink, a hand movement, a head turn, or the raising of an eyebrow can occur, coming to completion by the end of the word (Condon and Ogston 1967).

In addition to the synchrony among the speaker's intonational phrases, gestures and facial expressions, listeners may also tend to move in synchrony with the speaker (Condon and Ogston 1971). The boundary of the listener's gestures are coincident with the boundary of the speaker's gestures, although they often exhibit independent types of movements. This *interactional synchrony* occurs up to the word level (Kendon 1972). A subject not directly involved (i.e. one not looking directly at the speaker) might move in coordination with the speaker. Indeed, participants of a conversation react to the speaker's flow of speech (Kendon 1972).

Synchrony among body and facial motions occurs at all levels of speech, including the phoneme, the syllable, the word, the intonational phrase and the utterance (O'Connell and Slaymaker 1984, Dittman and Llewellyn 1967). That is, some movements are adapted to the phoneme level (like an eye blink), while others occur at the word level (like a frown) or even at the phrase level (like a hand gesture) (Condon and Ogston 1971). Facial expressions denoting emphasis generally match the intonationally emphasized segment. At a major pause or change of topic in conversation, a complete change of body posture may even occur.

7.1.2 The Link Between Facial Expressions and Speech

Facial expressions occur continuously during speech, both complementing and reinforcing the information delivered in the audio channel. For example, the raising of eyebrows may complement the corresponding speech by signaling surprise or may punctuate an emphatic (e.g. contrastive) element of the speech. Ekman (1989) differentiates between those facial expressions used as emotional signals and those used as conversational signals. The former (e.g., surprised expressions) are tied to emotion, while the latter (e.g., raised

eyebrows) are associated with intonation. Generally the conversational signals, which appear in concert with the emotional signals, are learned by observation, stereotyped and performed unconsciously by the speaker.

Facial expressions can be classified into the following groups, all of which must be modeled in order to obtain a complete, believable facial animation (Ekman 1989):

Emblems are produced to replace common verbal expressions and correspond to movements whose meanings are very well-known and culturally dependent (e.g. nodding instead of saying "yes"). Emblems can be directly translated into verbal statements and are produced consciously by the speaker. They may occur when words are blocked, when there is too much noise to talk, when people are too far from each other, or in order to iconically accentuate what is being said.

Emotion Emblems (also called referential expressions or mock expressions) are made to convey signals about emotions that are being referenced, but not currently being experienced, by the speaker. For example, it is quite common to wrinkle the nose when speaking about something "disgusting." Such referential movements are part of the corresponding emotional ones (e.g. wrinkling the nose is part of the facial expression of disgust).

Conversational Signals (also called illustrators) are made to accentuate or emphasize speech, and most often involve movement of the eyebrows. Ekman (1989) found raised eyebrows and frowns among the commonly used conversational signals. Head and eye motions can also illustrate a word, as when an accented word is accompanied by a rapid head movement (Hadar *et al.* 1983a, Bull and Connelly 1985), or when a blink occurs on a stressed vowel (Condon and Ogston 1971). When a conversational signal is used to accent a word, it is often termed a *baton*. When it stretches out over a syntactic portion of the sentence for emphasis, it is generally called an *underliner*.

Punctuators are movements which occur on pauses due to hesitation or to signal punctuation marks (such as a comma or an exclamation point). Certain distinct types of head movements occur during pauses. For example, a boundary point between intermediate phrases may be underlined by slow movement of the head, while a final

pause may coincide with stillness (Hadar *et al.* 1983a). Eye blinks and smiles can also occur during pauses (Condon and Ogston 1971). Some pauses, especially pauses filled with 'ah' or 'uh', serve as a signal of the speaker's searching for a word, and are frequently accompanied by frowns or raised brows with upward-looking eyes. Raising the eyebrows toward the end of an utterance often signals that something astonishing or unbelievable has been said. A frown, on the other hand, usually reflects a context of seriousness or anxiety. A question mark can be accompanied by movement of the eyebrows, particularly when the utterance's status as a question is based on intonation rather than syntax.

Regulators are movements that help the interaction between the speaker and the listener. Section 7.2.3 provides more detail about this group of facial actions.

Listener Responses are part of the regulator group, but involve the listener rather than the speaker. In any face-to-face conversation, the listener provides feedback to the speaker. A smile with or without raised eyebrows and/or a head nod can be a signal that the listener agrees with or at least understands what the speaker is trying to communicate. Conversely, raised eyebrows alone may signify the incredulity of the listener. A frown is often used to mark misunderstanding or incomprehension.

Manipulators correspond to biological requirements of the face, such as blinking the eyes to keep them moist or wetting the lips.

Affect Displays are the facial expressions of emotion.

7.1.3 Temporal Characteristics of Facial Actions

Facial movement can be described by the following three temporal parameters:

1. Onset duration: how long the expression takes to appear.
2. Apex duration: how long the expression remains in the apex position.
3. Offset duration: how long the expression takes to disappear.

Facial expressions of emotion differ in these parameters. For example, expressions of sadness have a slow offset, while expressions of happiness have a short onset. When an expression is used to deceive the listener (e.g. a polite smile vs. a smile of happiness), it often differs from the emotion expression in duration (apex), appearing either too late or too early. Although these parameters are extremely important for animation purposes, scant data on their values is available and few vision systems are currently able to extract the temporal parameters (Essa 1994, Yacoob and Davis 1994). Pelachaud *et al.* (1995) use these 3 parameters to specify a facial expression. Kalra *et al.* (1991) employ four parameters: attack (onset), decay, sustain (apex), and release (offset).

7.2 Eye Behavior

Visual behavior is an important feature whose main functions are to help regulate the flow of conversation, to signal the search for feedback during an interaction, to request information, to express emotion or to influence another person's behavior (Beattie 1981, Walker and Trimboli 1983, Webbink 1986). Humans are very sensitive to eye behavior and are able to perceive the slightest change in eye direction.

Eye contact is an important non-verbal method of establishing relationships and communicating with others. Eye contact tends to increase with the degree of intimacy and friendship between the speaker and the listener, but tends to decrease when a person is lying or experiencing difficulties in organizing his speech. People who like each other not only look at each other more often, but also tend to seek mutual gaze.

Eye movements can be defined by the direction of the gaze, the point or points of fixation, the percentage of eye contact over gaze avoidance, and the duration of eye contact (Argyle and Cook 1976). A common metric for eye behavior is "interest". That is, eyes glance at an object of interest for longer periods of time.

7.2.1 Eye Movements

The eyes are always in motion, usually with very saccadic movement (rapid changes of the point of fixation). When the eyes are forced to fix on one point, the visual field blurs

within three seconds. When examining an object, the eyes scan from the most salient features to the least in repeated cycles. For example, when looking at picture of a person, viewers are found to look primarily at the eyes (58% of the time), and then at the mouth (13%). The remaining regions of the face are scanned just 1% of the time each (Argyle and Cook 1976).

The change of focus (close focus vs. distant focus) is easily perceived by humans, but corresponds to only a six millimeter displacement of the irises. The irises are closer to each other during close focus than during distant focus.

7.2.2 The Eyes and the Environment

One's eyes are attracted by the objects in the surrounding environment. That is,

- Eyes are attracted by moving objects and tend to pursue them.
- Eyes fix on the object of a task. The length of eye-fixation depends on the change of state and the complexity of the task (Ullman 1984).
- Eyes localize the most important featured-objects and the nameable/recognizable objects (Ullman 1984).
- Eyes and hand movements are coordinated so that the eyes follow the last phase of hand manipulated objects (Ballard *et al.* 1992).

7.2.3 Gaze and Speech

Gaze also plays an important role in processing information during social interactions by synchronizing the conversation and determining when information is sought. Moreover, gaze is linked to intonational patterns and can be used to keep control of the communicative process (Argyle and Cook 1976). Eye contact is broken before one stops speaking, but is then reestablished before the other conversational participant begins speaking. The listener tends to look at the speaker (around 65% of the time) more than the speaker looks at the listener (less than 40% of the time) (Kendon 1967).

When replying to a question, the speaker breaks the gaze (more than 80% of the time). She generally looks at the listener more during fluent speech than during hesitant

speech, and has a tendency to look toward the end of a phrase. The direction of the gaze is maintained over the pause at the phrase boundary but often changes at the beginning of the next phrase. Such movements have a significance similar to punctuation marks in a written text.

Gaze behaviors such as those described above comprise part of a system of turn-taking that determines how people negotiate speaking turns in a conversation or any ritual meeting. Duncan (1974b) enumerates the different signals which comprise such a system as follows:[2]

- **Speaker-State-Signal**: displayed at the beginning of a speaking turn, and composed, at the least, of the speaker turning his head away from the listener and the starting of a speaker gesticulation.

- **Speaker-Within-Turn**: used when the speaker wants to keep his speaking turn and assure himself that the listener is following. In this phase, which occurs at the completion of a grammatical clause, the speaker turns his head toward the listener. This is frequently followed by the listener emitting a Listener-Backchannel signal to show her involvement in the conversation, which in turn may be followed by a Speaker-Continuation-Signal if the speaker wants to keep his turn.

- **Listener-Backchannel**: occurs generally after a Speaker-Within-Turn signal. When the speaker doesn't emit a signal by gazing at the listener, the listener can still emit a Listener-Backchannel signal, but the probability of this action by the listener varies with the action of the speaker (Duncan 1974b). In particular, it decreases if no signal has been emitted by the speaker. In this way the listener reacts to the behavior of the speaker. The signal is generally composed of a visual cue, which may be a head nod, a head shake, a gaze toward the speaker, or a smile, but may also include an audio cue. An audio cue may consist of completing the speaker's utterance, emitting a vocalization of agreement such as /mhm/, asking for clarification or restating the the speaker's utterance.

[2]These signals are also part of the regulator group of facial expressions introduced in Section 7.1.2.

- **Speaker-Continuation-Signal**: frequently follows a Speaker-Within-Turn signal. In such a case, the speaker turns his head (and eyes) away from the listener.

- **Speaker-Turn-Signal**: emitted when the speaker wants to give his speaking turn to the auditor. This signal may consist of several prosodic, gestural and syntactic clues. The speaker often turns his head toward the listener at the end of the utterance, assumes a more relaxed position and terminates all gesticulations of the hands and other body parts.

- **Listener-Request-Turn**: appears when the listener requests a speaking turn. The listener generally looks at the speaker. A smooth turn (i.e. an exchange of speaking turn without overlapping utterances) occurs when the speaker and listener emit a Speaker-Turn-Signal and a Listener-Request-Turn respectively.

7.2.4 Eye-Head Coordination

When people break eye contact to avoid gazing at one another, they usually move their heads to look away (Argyle 1975). A change in the direction of gaze is frequently accompanied by head movement (Argyle and Cook 1976, Boff and Lincoln 1988). For example, a sad person has a tendency not only to look downward with his eyes, but also to turn his head downward. In the case of a predictive event, the head generally starts moving before the eyes, which eventually follow with very rapid movements (Bizzi 1974).

Bizzi (1974) also found that a fast eye turn caused by an unexpected or sudden event is generally followed by a head turn 20 to 40 milliseconds later. The eyes move much faster than the head, stopping their movement when they have fixed on the target. When the head turns, the eyes rotate in the opposite direction in order to maintain focus on the target.

7.2.5 Blinking

The eyes blink frequently, serving not only to accentuate speech, but also to satisfy the biological need to keep them moist. In general, there is at least one blink of the eyes per utterance. These blinks can be characterized by the following parameters:

- the starting time of the eye closure
- the amount of time the eyes are closed
- the starting time of the opening of the eyes
- the amount time the eyes are completely open

Three forms of blinking occur:

- **Periodic blinks** correspond to the biological necessity of keeping the eyes wet. On average, they appear every 4.8 seconds (Argyle and Cook 1976), lasting about 1/4 sec., with 1/8 sec. of closure time, 1/24 sec. of closed eyes, and 1/12 sec. of aperture time (Grant 1969). The occurrence of blinks may also be emotion dependent. For example, when experiencing emotions such as fear, tension, anger, excitement or lying, the amount of blinking increases, while it tends to decrease during attentive or concentrated thought (Collier 1985).
- **Reflex blinks** occur when the cornea is irritated or when any sudden event (such as noise, bright light or a fast moving object approaching the eye) occurs.
- **Voluntary blinks** serve to emphasize speech and accentuate words.

7.3 Head Movements

Continuous sequences of head movements support the verbal stream. Head movements may be associated with emblems (nodding or shaking for agreement or disagreement), or with the maintenance of the flow of conversation (in a turn taking system). The direction of the head may be linked to an emotional state, such when a sad person turns her head downward, or may act as a deictic indicator.

Head movements also reflect encoding-decoding difficulties by coinciding with verbal hesitations and pauses within clauses. Hadar and his colleagues (1983a, 1983b) examined the relationship between such paralinguistic junctures and body movements. They

established a link between the temporal aspects of movement and prosodic features, distinguishing several classes of head motion based on amplitude and frequency. Slow movements were found to occur at 0.2 to 1.8 Hz, ordinary movements at 1.8 to 3.7 Hz, and rapid movements at 3.7 to 7.0 Hz. Postural Shifts were defined as linear movements (i.e. movements which change the axis of motion) of wide amplitude.

The results of Hadar's experiments show that distinct patterns of movement accompany linguistic features. For example, primary accents tend to be marked by rapid movements, while ordinary movements followed by stillness tend to indicate terminal points. Rapid movements occur just after dysfluencies, while postural shifts occur prior to the onset of speech after long pauses (Hadar *et al.* 1984). Rapid movement may also occur during speech dysfluencies, marked by repetition of syllables or words and short speech pauses within clauses. The occurrences of postural shifts at the beginning of speech, between speaking turns (Duncan 1973, 1974a, Wiemann and Knapp 1975) and at grammatical pauses (Kendon 1972), imply their involvement in speech production, regulation of turn-taking and identifying syntactic boundaries within clauses.

7.4 Implemented Systems

Pelachaud *et al.* (1995) examine how facial expressions, intonation and emotion are linked together. In an extension of their system, Cassell *et al.* (1994) consider the behavior of two agents in a dialogue situation using a ruled-based approach. More recently, Pelachaud and Prevost have applied similar techniques to the task of animating faces for automatically generated monologues, such as those described in Section 6.3. In both of these systems, head movements and facial expressions are characterized by their placement with respect to the linguistic utterance and their significance in transmitting information (Scherer 1980).

- *Syntactic* functions accompany the flow of speech and are synchronized at the verbal level. Facial movements (such as raising the eyebrows, nodding the head or blinking while saying "It costs NINE hundred dollars") can occur on an accented syllable (as a conversational signal) or on a pause (as a punctuator).

- *Dialogic* functions, such as gaze, regulate the flow of speech. Gaze behavior can be classified into four primary categories depending on its role in the conversation. *Planning* corresponds to the first phase of a turn when the speaker organizes her thoughts. The *comment* accompanies speech, occurring in parallel with accent and emphasis. *Control* regulates the exchange of speaking-turn. Finally, *feedback* is utilized to either collect or seek feedback from the listener.

A conversational signal (like the raising of an eyebrow) starts and ends with the accented word, while punctuator signals (such as smiling) coincide with pauses. Blinking (when occurring as a punctuator or a manipulator) is synchronized at the phoneme level. In this model, each facial expression is represented by two parameters: the *time of occurrence* and the *type*.

Gaze is generated using a sub-network, called *GAZE*, in a parallel transition network (PaT-Net, Becket 1994). PaT-Nets are a scheduling mechanism that are able to make decisions and execute actions. They allow rules that coordinate actions to be encoded as finite state automata that execute in parallel. Each of the four dialogic functions corresponds to a node which determines the appropriate actions to be performed. Each arc to a node corresponds to a condition for which the action should be performed. For each phoneme, the *GAZE* PaT-Net is entered and a transition is made on the node whose condition is true (see Figure 7.1). The different actions that can be performed include head turns, head shakes and head nods. Moreover, each action can be modulated depending on its function (e.g. head nods occurring as a *comment* or as *feedback* differ in amplitude). An example of the computation is presented in Section 7.5.

7.5 An Example

In this section we briefly describe how the system works for for the utterance in (114), which is produced by the monologue generator described in Section 6.3.

(114) "It costs NINE hundred dollars."

Here we are informally using capital letters to mark accented words. The utterance is decomposed into a sequence of phonemes with their durations and any corresponding

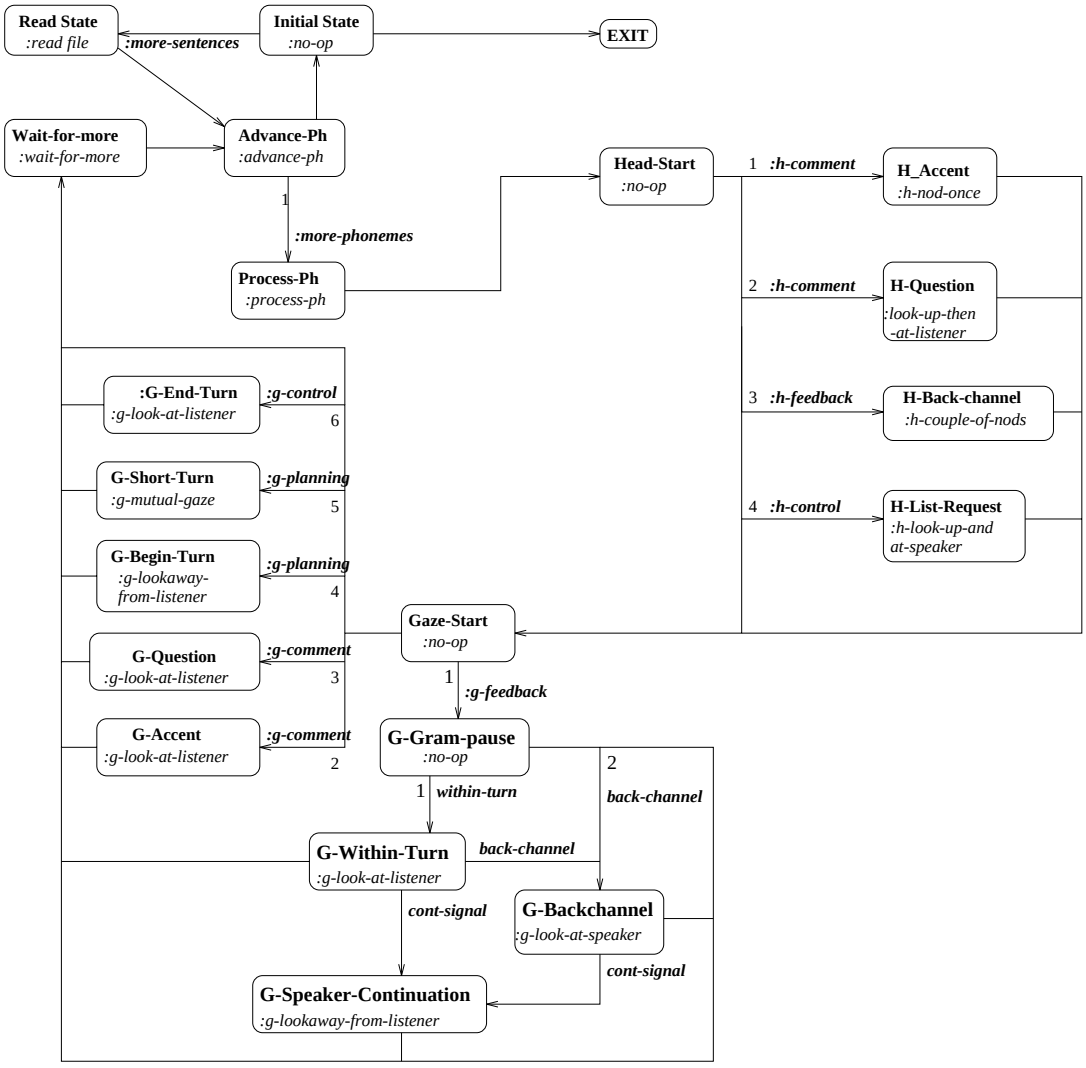


Figure 7.1: GAZE PaT-Net

accents specified, as shown in (115).

(115)
```
begin_intermediate
(accent-pi-L+H* IH <,0.090000> (accent-ph-L  TT <,0.060000>))
end_intermediate
begin_intermediate
_  <,0.020000>
KK <,0.080000> AO <,0.120000> SS <,0.070000> TT <,0.040000> SS <,0.080000>
NN <,0.050000> (accent-pi-H* AY <,0.170000> NN <,0.050000>
HH <,0.070000> AH <,0.070000> NN <,0.050000> DD <,0.040000>
               RR <,0.040000> IH <,0.060000> DD <,0.070000>
DD <,0.040000> AA <,0.130000> LL <,0.040000> ER <,0.090000> (accent-ph-L  ZZ <,0.060000>)
end_intermediate
```

In example (115), **H*** and **L+H*** correspond to a high pitch accent and a rising pitch accent respectively in Pierrehumbert's notation (see Chapter 3), while **L** corresponds to a low phrasal tone.

The algorithm first computes the lip shapes. For example, the program outputs the list of action units (**AU**s), each of which describes the movement of a muscle or muscle group, for each phonemic segment in the word *dollars*. The lip shapes associated with each segment and the effects of coarticulation rules are specified in (116) below.

(116)
```
Name: DD
name of au = AU12, intensity = 0.058333   /* backward coarticulation
name of au = AU28, intensity = 0.038889      rules: from IH of hundrEd
name of au = AU11, intensity = 0.097222      + some effects of AA's lip
name of au = AU20, intensity = 0.150000      shape */
name of au = AU25, intensity = 0.200000
Name: AA
name of au = AU11, intensity = 0.500000    /* extension of the lip +
name of au = AU20, intensity = 0.400000       small opening of the lip */
name of au = AU25, intensity = 0.600000
Name: LL
name of au = AU18, intensity = 0.180000    /* forward coarticulation
name of au = AU25, intensity = 0.200000       rules: from ER
name of au = AU11, intensity = 0.129630       + propagation of the lip
name of au = AU20, intensity = 0.103704       shape of AA (some
                                              lip extension remains) */
Name: ER
name of au = AU18, intensity = 0.720000     /* puckered lip shape */
name of au = AU25, intensity = 0.700000
Name: ZZ
name of au = rot, intensity = -0.045500     /* jaw rotation +
name of au = AU18, intensity = 0.180000        propagation of the lip
name of au = AU25, intensity = 0.200000        shape of ER (some puckered
                                               lip shape remains) */
```

Next, the program computes the facial expressions linked to the intonation as follows:

- A conversational signal occurs on the accented word "NINE", causing the eyebrows to raise. The movement starts on the initial phoneme /n/ and ends on the final /n/. There is also a blink of the eyes as part of the conversational signal. Similar movement on the word "amplifier" is illustrated in Figure 7.2.

- A punctuator signal is used to represent the end of the utterance, causing another blink to appear.

Finally, head movements are calculated as follows:

- At the beginning of the utterance, the head of the speaker moves away from the listener as part of the "planning" phase, where the speaker gathers his thoughts and prepares what he wants to say.

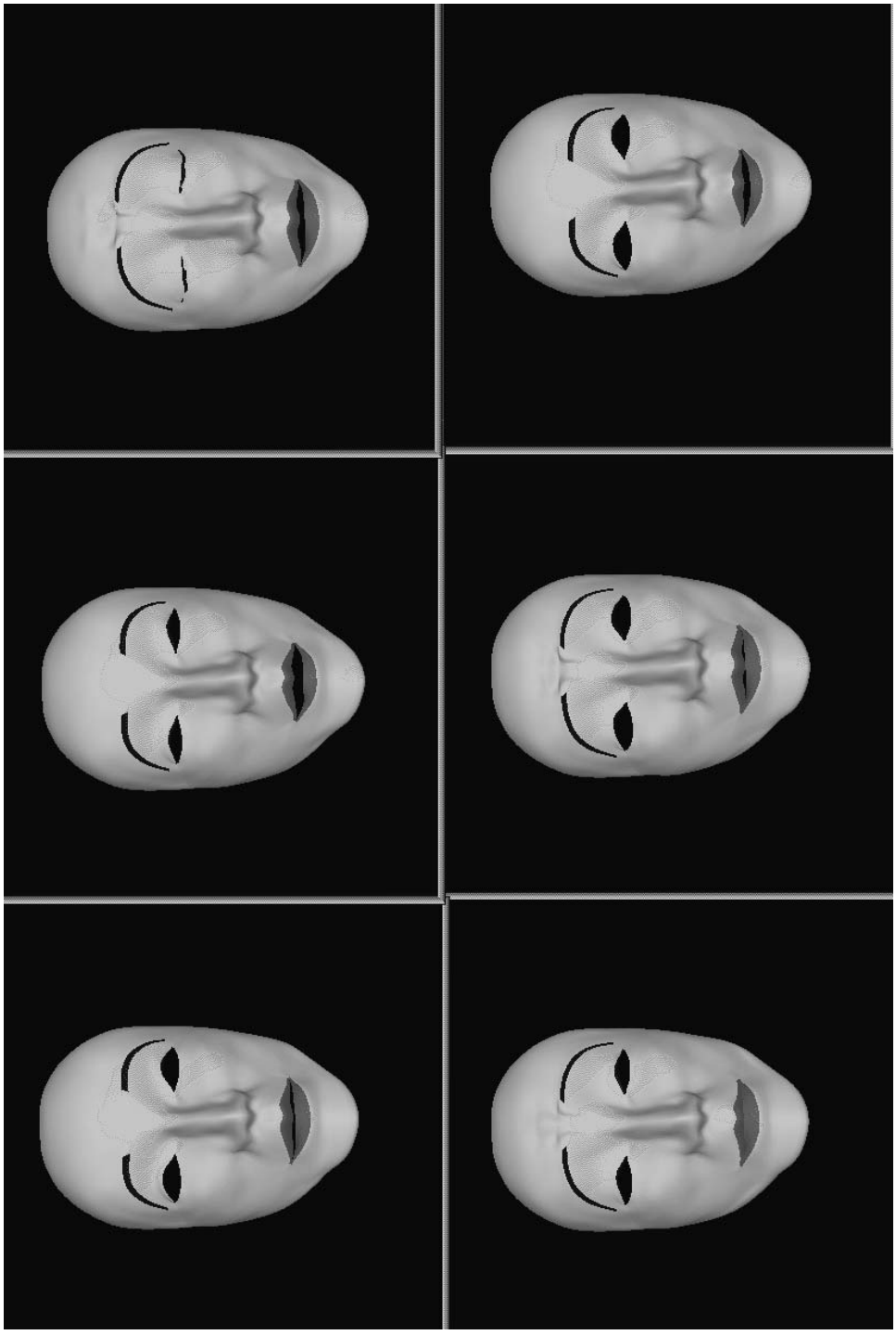


Figure 7.2: Facial Movements

- On the accented word (*nine*), the speaker gazes at the listener and punctuates the word with a head nod.

- To control the speaking turn, the speaker gazes at the listener at the end of the utterance, thereby signaling that he wants to give control of the conversation to the listener.

7.6 Summary

This chapter contains a brief summary of some of the related work on the meaning of facial expressions and their relation to the intonational theory and computational model described in the previous chapters. Section 7.4 provides a computational implementation that links the monologue generator described in Chapter 6 to the task of automatically producing contextually-appropriate facial animations. Facial expressions are an important communicative channel for study because, like intonation, they convey beliefs, intentions, attitudes and the types of semantic nuances encoded by the information structural representations presented throughout this dissertation. Our examination of how the synthesis of facial expressions and intonational features communicate a message in a specific context serves as a platform for testing hypotheses about the relationships and interactions between these two communicative channels.

Chapter 8

Conclusions

In this dissertation I have presented a model for the determination of intonation contours from context and provided an implemented system which applies this theory to the problem of generating spoken monologues with appropriate intonation from high-level semantic representations. Many earlier approaches at producing intonation from semantic representations have relied on *previous mention heuristics*, whereby pitch accents are assigned to lexical items based on whether or not they have previously occurred in the discourse. I demonstrate the inadequacy of this *previous mention* approach for handling a broad range of examples involving semantic contrasts, which require pitch accents to be allocated based on their ability to discriminate among available entities in the discourse model. Because the present model bases patterns of accentuation on sets of alternative entities in the database, I argue that it is better suited to the task of handling contrastive focus.

The present approach employs a two-level information structure representation which mediates between intonation and the discourse model. At the higher level, information structure divides an utterance into a theme, which provides the link to prior discourse, and a rheme, which provides the speaker's contribution to the material in the theme. At the intonational level, these delineations are congruent with intonational phrases. Within these top-level information structural constituents, the theme-focus and the rheme-focus mark the lexical items that bear pitch accents on the basis of their presentation of new information, or their ability to distinguish among alternatives in the discourse model.

By couching syntax, semantics, information structure and intonation in the framework of Combinatory Categorial Grammar, I present a model that treats all these levels of linguistic detail as structurally congruent. The adoption of this competence model for spoken language production simplifies the path from concept to speech by eliminating the need for separate prosodic and syntactic processes.

The major claims of this dissertation can be summarized as follows:

- Any model of intonation that assigns pitch accents based on the distinction between previously mentioned and "new" information does not account for accentual patterns that serve to distinguish among contrasting entities or properties in the discourse.

- Certain types of explicitly referential contrastive stress (i.e. accentual patterns that distinguish between two salient discourse entities) can be predicted by a model of *alternative sets* with respect to the discourse context.

- The *Information structure* representation effects a mapping between discourse context and intonation and provides a level of semantic detail that accounts for intonational phrasing, accentuation and discourse coherence.

- Intonation, information structure and syntax are structurally congruent.

- CCGs are instrumental for encoding the relationships among intonation, semantics, information structure and syntax.

- The problem of generating natural language is facilitated by a bi-level information structure representation. This representation associates the content planning phase with theme/rheme articulations and the sentence planning phase with focal distinctions within themes and rhemes.

- Information structure serves as a mechanism for incorporating several competing models of content planning (including schemata, rhetorical structure approaches and domain-driven approaches) into a hybrid framework for the high-level organization of text.

Chapter 2 introduced many of the disputed issues surrounding the structure, meaning and interpretation of intonational contours. This dissertation combines the theoretical work in phonology, semantics, syntax and discourse processing described at the beginning of the chapter with the body of applied research in artificial intelligence and speech synthesis. In Chapter 3, I presented an *information structure* formalism that mediates between intonation and discourse, and encodes the proper level of semantic information to account for both contextually bound accentuation patterns and intonational phrasing. Chapter 4 presents algorithms for processing discourse information and building the semantic and information structural representations necessary for generating context-appropriate intonation. The top-level division of an utterance into theme and rheme is calculated with respect to a collection of themes and rhemes from prior utterances, reminiscent of the calculation of C_b in centering theory. The lower-level articulations of theme-focus and rheme-focus are produced by an algorithm that considers both *previous mention* heuristics as well as alternative sets of properties and entities from the discourse model.

In Chapter 5, I expounded an intonational competence model in the framework of Combinatory Categorial Grammar, a mildly context-sensitive grammatical formalism which licenses congruent syntactic, prosodic and information structural constituents, and consequently represents a simplification over competence models that handle syntactic and prosodic aspects of speech separately. By constructing prosodic grammars to cover the range of intonational possibilities in descriptive monologues, I have demonstrated the appropriateness of the CCG/prosody model for spoken language production. This claim is further substantiated by the implementations described in Chapter 6, which serve as a platform for testing and refining the intonational competence model. A formal evaluation of the appropriateness of the speech produced by this model, however, is complicated by a number of factors which are also explicated in Chapter 6.

The generation model presented in Chapter 6 introduces a number of unique methods for organizing content and planning sentences. The information structural articulation of theme and rheme is incorporated into a hybrid content selection and organization approach that draws on both schemata and rhetorical structure theory. The lower-level

focus within information structural constituents is calculated at the sentence planning phase. This approach represents the first time the notion of information structure has been employed to ensure coherence for natural language generation.

In Chapter 7, I discussed related work concerning extra-linguistic modes of communication. This research, which focuses on the connection between intonation and facial expression, has produced a computational implementation that generates facial animation and speech from high-level semantic representations. Following the intonation generation programs described above, this implementation serves as a model for testing hypotheses about the relationship between facial expressions and speech.

The lack of consideration for contextual aspects of intonation in synthesized speech has often diminished the effectiveness of natural language interfaces. By generating spoken language from *concept* and *context*, the model which I have proffered makes a significant contribution to the resolution of this problem. In addition to addressing a variety of computational issues, this research indicates new directions for the study of contrast, the semantics of intonation and the production of prosody.

Appendix A

Sample Lexicon

```
%%  Lexicon Excerpt
%%
%% * stands for open circle in semantic representations
%% # stands for bullet in semantic representations


% Proper nouns
category(*stereofool, np(3,s):((*stereofool)^S)^S, nfocus, *stereofool).
category(#stereofool, np(3,s):((#stereofool)^S)^S, cfocus, #stereofool).
category(stereofool, np(3,s):(stereofool^S)^S, bg, stereofool).
category(*audiofad, np(3,s):((*audiofad)^S)^S, nfocus, *audiofad).
category(#audiofad, np(3,s):((#audiofad)^S)^S, cfocus, #audiofad).
category(audiofad, np(3,s):(audiofad^S)^S, bg, audiofad).
category(*it, np(3,s):((*it)^S)^S, nfocus, *it).
category(#it, np(3,s):((#it)^S)^S, cfocus, #it).
category(it, np(3,s):(it^S)^S, bg, it).


% Nouns
category(*x4, n(3,s):X^(*x4(X)), nfocus, *x4).
category(#x4, n(3,s):X^(#x4(X)), cfocus, #x4).
```

```
category(x4, n(3,s):X^(x4(X)), bg, x4).
category(*x5, n(3,s):X^(*x5(X)), nfocus, *x5).
category(#x5, n(3,s):X^(#x5(X)), cfocus, #x5).
category(x5, n(3,s):X^(x5(X)), bg, x5).
category(*amplifier, n(3,s):X^(*amplifier(X)), nfocus, *amplifier).
category(#amplifier, n(3,s):X^(#amplifier(X)), cfocus, #amplifier).
category(amplifier, n(3,s):X^(amplifier(X)), bg, amplifier).
category(*dollars, n(3,_):X^(*dollar_amount(X)), nfocus, *dollar_amount).
category(#dollars, n(3,_):X^(#dollar_amount(X)), cfocus, #dollar_amount).
category(dollars, n(3,_):X^(dollar_amount(X)), bg, dollar_amount).
category(*watts_per_channel, n(3,s):X^(*watts_per_channel(X)), nfocus,
         *watts_per_channel).
category(#watts_per_channel, n(3,s):X^(#watts_per_channel(X)), cfocus,
         #watts_per_channel).
category(watts_per_channel, n(3,s):X^(watts_per_channel(X)), bg,
         watts_per_channel).
category(*journal, n(3,s):X^(*journal(X)), nfocus, *journal).
category(#journal, n(3,s):X^(#journal(X)), cfocus, #journal).
category(journal, n(3,s):X^(journal(X)), bg, journal).


% Determiners
category(a, np(3,s):(X^S2)^indef(X, S1&S2)/n(3,s):(X^S1), bg, indef).
category(another, np(3,s):(X^S2)^indefrep(X, S1&S2)/n(3,s):(X^S1), bg,
         indefrep).
category(*another, np(3,s):(X^S2)^(*indefrep(X, S1&S2))/n(3,s):(X^S1),
         nfocus, *indefrep).
category(#another, np(3,s):(X^S2)^(#indefrep(X, S1&S2))/n(3,s):(X^S1),
         cfocus, #indefrep).
category(the, np(3,N):(X^S2)^def(X, S1&S2)/n(3,N):(X^S1), bg, def).
category(every, np(3,s):(X^S2)^forall(X, S1=>S2)/n(3,s):(X^S1), bg, forall).
category(*one, np(3,s):(X^S2)^amount(X, (S1&(*one(X)))&S2)/n(3,s):(X^S1),
         nfocus, *one).
category(#one, np(3,s):(X^S2)^amount(X, (S1&(#one(X)))&S2)/n(3,s):(X^S1),
```

```
        cfocus, #one).
category(one, np(3,s):(X^S2)^amount(X, (S1&one(X))&S2)/n(3,s):(X^S1),
        bg, one).
category(*two, np(3,s):(X^S2)^amount(X, (S1&(*two(X)))&S2)/n(3,s):(X^S1),
        nfocus, *two).
category(#two, np(3,s):(X^S2)^amount(X, (S1&(#two(X)))&S2)/n(3,s):(X^S1),
        cfocus, #two).
category(two, np(3,s):(X^S2)^amount(X, (S1&two(X))&S2)/n(3,s):(X^S1),
        bg, two).
category(*eight, np(3,s):(X^S2)^amount(X, (S1&(*eight(X)))&S2)/n(3,s):(X^S1),
        nfocus, *eight).
category(#eight, np(3,s):(X^S2)^amount(X, (S1&(#eight(X)))&S2)/n(3,s):(X^S1),
        cfocus, #eight).
category(eight, np(3,s):(X^S2)^amount(X, (S1&eight(X))&S2)/n(3,s):(X^S1),
        bg, eight).
category(*nine, np(3,s):(X^S2)^amount(X, (S1&(*nine(X)))&S2)/n(3,s):(X^S1),
        nfocus, *nine).
category(#nine, np(3,s):(X^S2)^amount(X, (S1&(#nine(X)))&S2)/n(3,s):(X^S1),
        cfocus, #nine).
category(nine, np(3,s):(X^S2)^amount(X, (S1&nine(X))&S2)/n(3,s):(X^S1),
        bg, nine).


% Adjectives
category(*solid_state, n(3,N):X^(Nn&(*solid_state(X)))/n(3,N):X^Nn, nfocus,
        *solid_state).
category(#solid_state, n(3,N):X^(Nn&(#solid_state(X)))/n(3,N):X^Nn, cfocus,
        #solid_state).
category(solid_state, n(3,N):X^(Nn&solid_state(X))/n(3,N):X^Nn, bg,
        solid_state).
category(*tube, n(3,N):X^(Nn&(*tube(X)))/n(3,N):X^Nn, nfocus, *tube).
category(#tube, n(3,N):X^(Nn&(#tube(X)))/n(3,N):X^Nn, cfocus, #tube).
category(tube, n(3,N):X^(Nn&tube(X))/n(3,N):X^Nn, bg, tube).
category(*audio, n(3,N):X^(Nn&(*audio(X)))/n(3,N):X^Nn, nfocus, *audio).
```

```
category(#audio, n(3,N):X^(Nn&(#audio(X)))/n(3,N):X^Nn, cfocus, #audio).
category(audio, n(3,N):X^(Nn&audio(X))/n(3,N):X^Nn, bg, audio).
category(*stereo, n(3,N):X^(Nn&(*stereo(X)))/n(3,N):X^Nn, nfocus, *stereo).
category(#stereo, n(3,N):X^(Nn&(#stereo(X)))/n(3,N):X^Nn, cfocus, #stereo).
category(stereo, n(3,N):X^(Nn&stereo(X))/n(3,N):X^Nn, bg, stereo).
category(*another, n(3,s):X^(Nn&(*another(X)))/n(3,s):X^Nn, nfocus, *another).
category(#another, n(3,s):X^(Nn&(#another(X)))/n(3,s):X^Nn, cfocus, #another).
category(another, n(3,s):X^(Nn&another(X))/n(3,s):X^Nn, bg, another).


% Transitive verbs
category(*(is), (s:(act^pres)^(*isa(X,Y))\np(3,s):X)/np(_,_):Y, nfocus, *isa).
category(#(is), (s:(act^pres)^(#isa(X,Y))\np(3,s):X)/np(_,_):Y, cfocus, #isa).
category(is, (s:(act^pres)^isa(X,Y)\np(3,s):X)/np(_,_):Y, bg, isa).
category(*are, (s:(act^pres)^(*isa(X,Y))\np(_,p):X)/np(_,_):Y, nfocus, *isa).
category(#are, (s:(act^pres)^(#isa(X,Y))\np(_,p):X)/np(_,_):Y, cfocus, #isa).
category(are, (s:(act^pres)^isa(X,Y)\np(_,p):X)/np(_,_):Y, bg, isa).
category(*was, (s:(act^past)^(*isa(X,Y))\np(_,s):X)/np(_,_):Y, nfocus, *isa).
category(#was, (s:(act^past)^(#isa(X,Y))\np(_,s):X)/np(_,_):Y, cfocus, #isa).
category(was, (s:(act^past)^isa(X,Y)\np(_,s):X)/np(_,_):Y, bg, isa).
category(*were, (s:(act^past)^(*isa(X,Y))\np(_,s):X)/np(_,_):Y, nfocus, *isa).
category(#were, (s:(act^past)^(#isa(X,Y))\np(_,s):X)/np(_,_):Y, cfocus, #isa).
category(were, (s:(act^past)^isa(X,Y)\np(_,s):X)/np(_,_):Y, bg, isa).

category(*costs, (s:(act^pres)^(*cost(X,Y))\np(3,s):X)/np(_,_):Y, nfocus,
        *cost).
category(#costs, (s:(act^pres)^(#cost(X,Y))\np(3,s):X)/np(_,_):Y, cfocus,
        #cost).
category(costs, (s:(act^pres)^cost(X,Y)\np(3,s):X)/np(_,_):Y, bg, cost).
category(*cost, (s:(act^pres)^(*cost(X,Y))\np(_,p):X)/np(_,_):Y, nfocus,
        *cost).
category(#cost, (s:(act^pres)^(#cost(X,Y))\np(_,p):X)/np(_,_):Y, cfocus,
        #cost).
category(cost, (s:(act^pres)^cost(X,Y)\np(_,p):X)/np(_,_):Y, bg, cost).
```

```
category(*cost, (s:(act^past)^(*cost(X,Y))\np(_,_):X)/np(_,_):Y, nfocus,
        *cost).
category(#cost, (s:(act^past)^(#cost(X,Y))\np(_,_):X)/np(_,_):Y, cfocus,
        #cost).
category(cost, (s:(act^past)^cost(X,Y)\np(_,_):X)/np(_,_):Y, bg, cost).

category(*produces, (s:(act^pres)^(*produce(X,Y))\np(3,s):X)/np(_,_):Y,
        nfocus, *produce).
category(#produces, (s:(act^pres)^(#produce(X,Y))\np(3,s):X)/np(_,_):Y,
        cfocus, #produce).
category(produces, (s:(act^pres)^produce(X,Y)\np(3,s):X)/np(_,_):Y,
        bg, produce).
category(*produce, (s:(act^pres)^(*produce(X,Y))\np(_,p):X)/np(_,_):Y,
        nfocus, *produce).
category(#produce, (s:(act^pres)^(#produce(X,Y))\np(_,p):X)/np(_,_):Y,
        cfocus, #produce).
category(produce, (s:(act^pres)^produce(X,Y)\np(_,p):X)/np(_,_):Y,
        bg, produce).
category(*produced, (s:(act^past)^(*produce(X,Y))\np(_,_):X)/np(_,_):Y,
        nfocus, *produce).
category(#produced, (s:(act^past)^(#produce(X,Y))\np(_,_):X)/np(_,_):Y,
        cfocus, #produce).
category(produced, (s:(act^past)^produce(X,Y)\np(_,_):X)/np(_,_):Y,
        bg, produce).

category(*praises, (s:(act^pres)^(*praise(X,Y))\np(3,s):X)/np(_,_):Y,
        nfocus, *praise).
category(#praises, (s:(act^pres)^(#praise(X,Y))\np(3,s):X)/np(_,_):Y,
        cfocus, #praise).
category(praises, (s:(act^pres)^praise(X,Y)\np(3,s):X)/np(_,_):Y,
        bg, praise).
category(*praise, (s:(act^pres)^(*praise(X,Y))\np(_,p):X)/np(_,_):Y,
        nfocus, *praise).
category(#praise, (s:(act^pres)^(#praise(X,Y))\np(_,p):X)/np(_,_):Y,
```

```
        cfocus, #praise).
category(praise, (s:(act^pres)^praise(X,Y)\np(_,p):X)/np(_,_):Y,
        bg, praise).
category(*praised, (s:(act^past)^(*praise(X,Y))\np(_,_):X)/np(_,_):Y,
        nfocus, *praise).
category(#praised, (s:(act^past)^(#praise(X,Y))\np(_,_):X)/np(_,_):Y,
        cfocus, #praise).
category(praised, (s:(act^past)^praise(X,Y)\np(_,_):X)/np(_,_):Y,
        bg, praise).
category(*praised, vp:(pas^pres)^Y^(*praise(X,Y))/pp:X^by(X), nfocus,
        *praise).
category(#praised, vp:(pas^pres)^Y^(#praise(X,Y))/pp:X^by(X), cfocus,
        #praise).
category(praised, vp:(pas^pres)^Y^praise(X,Y)/pp:X^by(X), bg, praise).
category(*praised, vp:(pas^past)^Y^(*praise(X,Y))/pp:X^by(X), nfocus,
        *praise).
category(#praised, vp:(pas^past)^Y^(#praise(X,Y))/pp:X^by(X), cfocus,
        #praise).
category(praised, vp:(pas^past)^Y^praise(X,Y)/pp:X^by(X), bg, praise).

category(*reviles, (s:(act^pres)^(*revile(X,Y))\np(3,s):X)/np(_,_):Y,
        nfocus, *revile).
category(#reviles, (s:(act^pres)^(#revile(X,Y))\np(3,s):X)/np(_,_):Y,
        cfocus, #revile).
category(reviles, (s:(act^pres)^revile(X,Y)\np(3,s):X)/np(_,_):Y,
        bg, revile).
category(*revile, (s:(act^pres)^(*revile(X,Y))\np(_,p):X)/np(_,_):Y,
        nfocus, *revile).
category(#revile, (s:(act^pres)^(#revile(X,Y))\np(_,p):X)/np(_,_):Y,
        cfocus, #revile).
category(revile, (s:(act^pres)^revile(X,Y)\np(_,p):X)/np(_,_):Y,
        bg, revile).
category(*reviled, (s:(act^past)^(*revile(X,Y))\np(_,_):X)/np(_,_):Y,
        nfocus, *revile).
```

```
category(#reviled, (s:(act^past)^(#revile(X,Y))\np(_,_):X)/np(_,_):Y,
         cfocus, #revile).
category(reviled, (s:(act^past)^revile(X,Y)\np(_,_):X)/np(_,_):Y,
         bg, revile).
category(*reviled, vp:(pas^pres)^Y^(*revile(X,Y))/pp:X^by(X),
         nfocus, *revile).
category(#reviled, vp:(pas^pres)^Y^(#revile(X,Y))/pp:X^by(X),
         cfocus, #revile).
category(reviled, vp:(pas^pres)^Y^revile(X,Y)/pp:X^by(X),
         bg, revile).
category(*reviled, vp:(pas^past)^Y^(*revile(X,Y))/pp:X^by(X),
         nfocus, *revile).
category(#reviled, vp:(pas^past)^Y^(#revile(X,Y))/pp:X^by(X),
         cfocus, #revile).
category(reviled, vp:(pas^past)^Y^revile(X,Y)/pp:X^by(X),
         bg, revile).


% Auxiliary verbs
category(am, (s:(act^presp)^S\np(1,s):E)/vp:(act^presp)^(E^S), bg, presp).
category(are, (s:(act^presp)^S\np(2,s):E)/vp:(act^presp)^(E^S), bg, presp).
category(is, (s:(act^presp)^S\np(3,s):E)/vp:(act^presp)^(E^S), bg, presp).
category(are, (s:(act^presp)^S\np(_,p):E)/vp:(act^presp)^(E^S), bg, presp).
category(have, (s:(act^pastp)^S\np(1,s):E)/vp:(act^pastp)^(E^S), bg, pastp).
category(have, (s:(act^pastp)^S\np(2,s):E)/vp:(act^pastp)^(E^S), bg, pastp).
category(has, (s:(act^pastp)^S\np(3,s):E)/vp:(act^pastp)^(E^S), bg, pastp).
category(have, (s:(act^pastp)^S\np(_,p):E)/vp:(act^pastp)^(E^S), bg, pastp).
category(was, (s:(pas^past)^S\np(1,s):X)/(vp:(pas^past)^X^S), bg, pas).
category(was, ((s:(pas^past)^S\np(1,s):X)/NPorPP)/(vp:(pas^past)^X^S/NPorPP),
         bg, pas).
category(were, (s:(pas^past)^S\np(2,s):X)/(vp:(pas^past)^X^S), bg, pas).
category(were, ((s:(pas^past)^S\np(2,s):X)/NPorPP)/(vp:(pas^past)^X^S/NPorPP),
         bg, pas).
category(was, (s:(pas^past)^S\np(3,s):X)/(vp:(pas^past)^X^S), bg, pas).
```

```
category(was, ((s:(pas^past)^S\np(3,s):X)/NPorPP)/(vp:(pas^past)^X^S/NPorPP),
        bg, pas).
category(were, (s:(pas^past)^S\np(_,p):X)/(vp:(pas^past)^X^S), bg, pas).
category(were, ((s:(pas^past)^S\np(_,p):X)/NPorPP)/(vp:(pas^past)^X^S/NPorPP),
        bg, pas).
category(am, (s:(pas^pres)^S\np(1,s):X)/(vp:(pas^pres)^X^S), bg, pas).
category(am, ((s:(pas^pres)^S\np(1,s):X)/NPorPP)/(vp:(pas^pres)^X^S/NPorPP),
        bg, pas).
category(are, (s:(pas^pres)^S\np(2,s):X)/(vp:(pas^pres)^X^S), bg, pas).
category(are, ((s:(pas^pres)^S\np(2,s):X)/NPorPP)/(vp:(pas^pres)^X^S/NPorPP),
        bg, pas).
category(is, (s:(pas^pres)^S\np(3,s):X)/(vp:(pas^pres)^X^S), bg, pas).
category(is, ((s:(pas^pres)^S\np(3,s):X)/NPorPP)/(vp:(pas^pres)^X^S/NPorPP),
        bg, pas).
category(are, (s:(pas^pres)^S\np(_,p):X)/(vp:(pas^pres)^X^S), bg, pas).
category(are, ((s:(pas^pres)^S\np(_,p):X)/NPorPP)/(vp:(pas^pres)^X^S/NPorPP),
        bg, pas).
category(does, (s:(act^presnf)^S\np(3,s):E)/vp:(act^nfin)^(E^S), bg, presnf).
%category(does, (s:(act^presnf)^S/vp:(act^nfin)^(E^S))/np(3,s):E, bg, presnf).
category(do, (s:(act^presnf)^S\np(_,p):E)/vp:(act^nfin)^(E^S), bg, presnf).
%category(do, (s:(act^presnf)^S/vp:(act^nfin)^(E^S))/np(_,p):E, bg, presnf).
category(did, (s:(act^pastnf)^S\np(_,_):E)/vp:(act^nfin)^(E^S), bg, pastnf).
%category(did, (s:(act^pastnf)^S/vp:(act^nfin)^(E^S))/np(_,_):E, bg, pastnf).


% Prepositions
category(by, pp:(X^S1)^S2^by(X)/np(_,_):(X^S1)^S2, bg, pas).
category(by, pp:(X^S1)^S2^by(X)/np(_,_):(X^S1)^S2, bg, by).
category(to, pp:(X^S1)^S2^to(X)/np(_,_):(X^S1)^S2, bg, pas).
category(to, pp:(X^S1)^S2^to(X)/np(_,_):(X^S1)^S2, bg, give).
```

Appendix B

Sample Output: Monologue Generator

```
| ?- generate(intention(believe(h1,good_to_buy(e1))), S1),
     generate(intention(believe(h1,good_to_buy(e2))), S2).

RH_DE_TREE:

de_tree(
   props(e1,
      [holds(defn(isa(e1,amplifier))),
       holds(design(e1,solid_state),pres),
       holds(cost(e1,e9),pres),
       holds(produce(e1,e7),pres),
       holds(but(praise(e4,e1),revile(e5,e1)),past)]),
   supports(e1,
      [de_tree(
          props(e9,
             [holds(isa(e9,dollar_amount),pres),
              holds(amount(e9,800),pres)]),
          supports(e9,[])),
       de_tree(
```

```
          props(e7,
             [holds(isa(e7,watts_per_channel),pres),
              holds(amount(e7,100),pres)]),
           supports(e7,[])),
       de_tree(
          props(e4,
             [holds(isa(e4,journal),pres),
              holds(subject(e4,stereo),pres)]),
          supports(e4,[])),
       de_tree(
          props(e5,
             [holds(isa(e5,journal),pres),
              holds(subject(e5,stereo),pres)]),
          supports(e5,[]))]))


IS_TREES:

[node(e1,
   nouns([]),
   adjs([]),
   clauses(
      [infostruc(
         sem(defn(isa(e1,c1))),
         th(e1),
         rh(x^isa(x,c1)))]),
   sub(
      [[node(c1,
           nouns([isa(c1,amplifier)]),
           adjs([design(c1,solid_state)]),
           clauses([]),sub([]))]])),
 node(e1,
   nouns([]),
   adjs([]),
```

```
  clauses(
     [infostruc(
        sem(cost(e1,e9)),
        th(e1),
        rh(x^cost(x,e9))),
      infostruc(
        sem(and(produce(e1,e7))),
        th(e1),
        rh(x^and(produce(x,e7))))]),
  sub(
     [[node(e9,
         nouns([isa(e9,dollar_amount)]),
         adjs([amount(e9,compound(eight,hundred))]),
         clauses([]),sub([]))],
      [node(e7,
         nouns([isa(e7,watts_per_channel)]),
         adjs([amount(e7,compound(one,hundred))]),
         clauses([]),sub([]))]])),
node(e1,
  nouns([]),
  adjs([]),
  clauses(
     [infostruc(
        sem(praise(e4,e1)),
        th(e1),
        rh(x^praise(e4,x))),
      infostruc(
        sem(but(revile(e5,e1))),
        th(e1),rh(x^but(revile(e5,x))))]),
  sub(
     [[node(e4,
         nouns([isa(e4,journal)]),
         adjs([subject(e4,stereo)]),
         clauses([]),sub([]))],
```

```
       [node(e5,
          nouns([isa(e5,journal)]),
          adjs([subject(e5,stereo)]),clauses([]),sub([]))]]))]

PLAN SENTENCE

IS:
[infostruc(sem(defn(isa(e1,c1))),th(e1),rh(x^isa(x,c1)))]

CONTRASTIVE FOCUS:
node(e1,
   nouns([]),
   adjs([]),
   clauses(
      [infostruc(
         sem(defn(isa(*e1,*c1))),
         th(*e1),
         rh(x^isa(x,*c1)))]),
   sub(
      [[node(c1,
         nouns([isa(c1,*amplifier)]),
         adjs([*design(c1,*solid_state)]),
         clauses([]),sub([]))]]))

SEMANTICS:
np(3,s): (e1^_85807)^def(e1,*x4(e1)&_85807) @ u/rh(_)

s: (act^pres)^indef(c1,(*amplifier(c1)& *solid_state(c1))&
                        isa(e1,c1))\np(3,s):e1 @ rh(_)
```

```
OUTPUT: the x4@lhstar @lh

OUTPUT: is a solid_state@hstar amplifier@hstar @lls



PLAN SENTENCE

IS:
[infostruc(sem(cost(e1,e9)),th(e1),rh(x^cost(x,e9))),
 infostruc(sem(and(produce(e1,e7))),th(e1),rh(x^and(produce(x,e7))))]

CONTRASTIVE FOCUS:
node(e1,
  nouns([]),
  adjs([]),
  clauses(
     [infostruc(
        sem(*cost(e1,*e9)),
        th(e1),rh(x^ (*cost(x,*e9)))),
      infostruc(
        sem(and(*produce(e1,*e7))),
        th(e1),
        rh(x^and(*produce(x,*e7))))]),
  sub(
     [[node(e9,
         nouns([isa(e9,*dollar_amount)]),
         adjs([*amount(e9,compound(*eight,*hundred))]),
         clauses([]),sub([]))],
      [node(e7,
         nouns([isa(e7,*watts_per_channel)]),
         adjs([amount(e7,compound(*one,hundred))]),
         clauses([]),sub([]))]]))

SEMANTICS:
```

```
np(3,s): (it^_90785)^_90785 @ u/rh(_)

s: (act^pres)^amount(e9,((*dollar_amount(e9)& *hundred(e9))& *eight(e9))
                        & *cost(it,e9))\np(3,s):it @ rh(_)/rh(_)

s: (act^pres)^amount(e7,((*watts_per_channel(e7)&hundred(e7))& *one(e7))
                        & and(*produce(it,e7)))\np(3,s):it @ rh(_)

OUTPUT: it

OUTPUT: costs@hstar eight@hstar hundred@hstar dollars@hstar @lhb

OUTPUT: and produces@hstar one@hstar hundred watts_per_channel@hstar @lls


PLAN SENTENCE

IS:
[infostruc(sem(praise(e4,e1)),th(e1),rh(x^praise(e4,x))),
 infostruc(sem(but(revile(e5,e1))),th(e1),rh(x^but(revile(e5,x))))]

CONTRASTIVE FOCUS:
node(e1,
  nouns([]),
  adjs([]),
  clauses(
     [infostruc(
        sem(#praise(#e4,e1)),
        th(e1),
        rh(x^(#praise(#e4,x)))),
      infostruc(
        sem(but(#revile(#e5,e1))),
        th(e1),rh(x^but(#revile(#e5,x))))]),
  sub(
```

```
     [[node(e4,
         nouns([isa(e4,*journal)]),
         adjs([*subject(e4,*stereo)]),
         clauses([]),sub([]))],
       [node(e5,
         nouns([isa(e5,journal)]),
         adjs([subject(e5,stereo)]),
         clauses([]),sub([]))]]))

SEMANTICS:
np(3,s): (it^_98931)^_98931 @ u/th(_)

s: (pas^past)^ (#praise(#stereofool,it))\np(3,s):it @ rh(_)/drh(_)
np(3,s): (e4^_99474)^indef(e4,(*journal(e4)& *stereo(e4))&_99474) @ rh(_)/rh(_)

s: (pas^past)^but(#revile(#audiofad,it))\np(3,s):it @ rh(_)/drh(_)
np(3,s): (e5^_100248)^(*indefrep(e5,(journal(e5)&stereo(e5))
                                    &_100248)) @ rh(_)/rh(_)



OUTPUT: it

OUTPUT: was praised@hhash by stereofool@dhhash @lhb

OUTPUT: a stereo@hstar journal@hstar @lhb

OUTPUT: but was reviled@hhash by audiofad@dhhash @lhb

OUTPUT: another@hstar stereo journal @lls
```

```
RH_DE_TREE:


de_tree(
    props(e2,
       [holds(design(e2,tube),pres),
        holds(cost(e2,e10),pres),
        holds(produce(e2,e8),pres),
        holds(praise(conj(e4,e5),e2),past)]),
    supports(e2,
       [de_tree(
          props(e10,
             [holds(isa(e10,dollar_amount),pres),
              holds(amount(e10,900),pres)]),
          supports(e10,[])),
        de_tree(
          props(e8,
             [holds(isa(e8,watts_per_channel),pres),
              holds(amount(e8,200),pres)]),
          supports(e8,[])),
        de_tree(
          props(e4,[]),
          supports(e4,[])),
        de_tree(
          props(e5,[]),
          supports(e5,[]))]))



IS_TREES:

[node(e2,
   nouns([]),
   adjs([]),
```

```
  clauses(
     [infostruc(
        sem(defn(isa(e2,c2))),
        th(e2),
        rh(x^isa(x,c2)))]),
  sub(
     [[node(c2,
         nouns([isa(c2,amplifier)]),
         adjs([design(c2,tube)]),
         clauses([]),
         sub([]))]])),
node(e2,
  nouns([]),
  adjs([]),
  clauses(
     [infostruc(
        sem(cost(e2,e10)),
        th(e2),
        rh(x^cost(x,e10))),
      infostruc(
        sem(and(produce(e2,e8))),
        th(e2),
        rh(x^and(produce(x,e8))))]),
  sub(
     [[node(e10,
         nouns([isa(e10,dollar_amount)]),
         adjs([amount(e10,compound(nine,hundred))]),
         clauses([]),
         sub([]))],
      [node(e8,
         nouns([isa(e8,watts_per_channel)]),
         adjs([amount(e8,compound(two,hundred))]),
         clauses([]),
         sub([]))]])),
```

```
 node(e2,
   nouns([]),
   adjs([]),
   clauses(
      [infostruc(
         sem(praise(conj(e4,e5),e2)),
         th(e2),
         rh(x^praise(conj(e4,e5),x)))]),
   sub([[],[]]))]
```

```
PLAN SENTENCE

IS:
[infostruc(sem(defn(isa(e2,c2))),th(e2),rh(x^isa(x,c2)))]

CONTRASTIVE FOCUS:
node(e2,
   nouns([]),
   adjs([]),
   clauses(
      [infostruc(
         sem(defn(isa(#e2,#c2))),
         th(#e2),
         rh(x^isa(x,#c2)))]),
   sub(
      [[node(c2,
         nouns([isa(c2,amplifier)]),
         adjs([design(c2,*tube)]),
         clauses([]),
         sub([]))]]))

SEMANTICS:
```

```
np(3,s): (e2^_143595)^def(e2,#x5(e2)&_143595) @ u/rh(_)

s: (act^pres)^indef(c2,(amplifier(c2)& #tube(c2))&
                        isa(e2,c2))\np(3,s):e2 @ rh(_)


OUTPUT: the x5@lhhash @lh

OUTPUT: is a tube@hhash amplifier @lls


PLAN SENTENCE

IS:
[infostruc(sem(cost(e2,e10)),th(e2),rh(x^cost(x,e10))),
 infostruc(sem(and(produce(e2,e8))),th(e2),rh(x^and(produce(x,e8))))]

CONTRASTIVE FOCUS:
node(e2,
   nouns([]),
   adjs([]),
   clauses(
      [infostruc(
         sem(cost(#e2,#e10)),
         th(#e2),
         rh(x^cost(x,#e10))),
       infostruc(
         sem(and(produce(#e2,#e8))),
         th(#e2),
         rh(x^and(produce(x,#e8))))]),
   sub(
      [[node(e10,
          nouns([isa(e10,dollar_amount)]),
          adjs([amount(e10,compound(*nine,hundred))]),
```

```
          clauses([]),
          sub([]))],
       [node(e8,
          nouns([isa(e8,watts_per_channel)]),
          adjs([amount(e8,compound(*two,hundred))]),
          clauses([]),
          sub([]))]]))

SEMANTICS:
np(3,s): ((#it)^_151490)^_151490@theme

s: (act^pres)^amount(e10,((dollar_amount(e10)&hundred(e10))& #nine(e10))
                         & cost(#it,e10))\np(3,s): (#it) @ rh(_)

s: (act^pres)^amount(e8,((watts_per_channel(e8)&hundred(e8))& #two(e8))
                         & and(produce(#it,e8)))\np(3,s): (#it) & rh(_)


OUTPUT: it@lhhash @lh

OUTPUT: costs nine@hhash hundred dollars @lhb

OUTPUT: and produces two@hhash hundred watts_per_channel @lls


PLAN SENTENCE

IS:
[infostruc(sem(praise(conj(e4,e5),e2)),th(e2),rh(x^praise(conj(e4,e5),x)))]

CONTRASTIVE FOCUS:
node(e2,
   nouns([]),
   adjs([]),
```

```
   clauses(
      [infostruc(
         sem(praise(#conj(e4,e5),#e2)),
         th(#e2),
         rh(x^praise(#conj(e4,e5),x)))]),
   sub([[],[]]))

SEMANTICS:
np(3,s): ((#it)^_155665)^_155665 & u/rh(_)

s: (pas^past)^praise(#conj(stereofool,audiofad),#it)\np(3,s): (#it) @ rh(_)



OUTPUT: it@lhhash @lh

OUTPUT: was praised by stereofool and@hhash audiofad @lls


S1 = [[[the,x4@lhstar,''@lh],
       [is,a,solid_state@hstar,amplifier@hstar,''@lls]],
      [[it],
       [costs@hstar,eight@hstar,hundred@hstar,dollars@hstar,''@lhb],
       [and,produces@hstar,one@hstar,hundred,watts_per_channel@hstar,''@lls]],
      [[it],
       [was,praised@hhash,by,stereofool@dhhash,''@lhb],
       [a,stereo@hstar,journal@hstar,''@lhb],
       [but,was,reviled@hhash,by,audiofad@dhhash,''@lhb],
       [another@hstar,stereo,journal,''@lls]]],

S2 = [[[the,x5@lhhash,''@lh],
       [is,a,tube@hhash,amplifier,''@lls]],
      [[it@lhhash,''@lh],
       [costs,nine@hhash,hundred,dollars,''@lhb],
```

```
 [and,produces,two@hhash,hundred,watts_per_channel,''@lls]],
[[it@lhhash,''@lh],
 [was,praised,by,stereofool,and@hhash,audiofad,''@lls]]]
```

Bibliography

Ajdukiewicz, K. (1935). Die syntaktische konnexitat. *Studia Philosophica*, 1:1–27. English translation in Storrs McCall (ed.), *Polish Logic 1920–1939*, Oxford University Press, pages 207–231.

Allen, J., Hunnicutt, S., and Klatt, D. (1987). *From Text to Speech: the MITalk System.* Cambridge University Press, Cambridge.

Argyle, M. (1975). *Bodily Communication.* Methuen and Co. Ltd, London.

Argyle, M. and Cook, M. (1976). *Gaze and Mutual gaze.* Cambridge University Press.

Ballard, D. (1992). Hand-eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London*, 337:331–339.

Bar-Hillel, Y. (1953). A quasi-arithmetical notation for syntactic description. *Language*, 29:47–58.

Beattie, G. (1981). Sequential temporal patterns of speech and gaze in dialogue. In Sebeok, T. and Umiker-Sebeok, J., editors, *Nonverbal Communication, Interaction, and Gesture*, pages 297–320. The Hague, New-York.

Becket, W. M. (1994). The Jack Lisp API. Technical Report MS-CIS-94-01, Graphics Lab 59, University of Pennsylvania.

Beckman, M. and Hirschberg, J. (1994). The ToBI annotation conventions. Manuscript, Ohio State University.

Beckman, M. and Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3:255–310.

Berman, A. and Szamosi, M. (1972). Observations on sentential stress. *Language*, 48:304–325.

Bird, S. (1991). Focus and phrasing in Unification Categorial Grammar. In Bird, S., editor, *Declarative Perspectives on Phonology*, pages 139–166. University of Edinburgh.

Bizzi, E. (1974). The coordination of eye-head movement. *Scientific American*, 531(4):100–106.

Boff, K. and Lincoln, J. (1988). *Human Perception and Performance*. Engineering data Compendum, Aerospace Medical Research Laboratory.

Bolinger, D. (1961). Contrastive accent and contrastive stress. *Language*, 37:83–96.

Bolinger, D. (1972). Accent is predictable (if you're a mind reader). *Language*, 48:633–644.

Bolinger, D. (1989). *Intonation and Its Uses*. Stanford University Press.

Brennan, S., Friedman, M., and Pollard, C. (1987). A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162, Stanford.

Bresnan, J. (1971). Sentence stress and syntactic transformations. *Language*, 47:257–280.

Bresnan, J., Kaplan, R., Peters, S., and Zaenen, A. (1982). Cross-serial dependencies in Dutch. *Linguistic Inquiry*, 13:613–36.

Bull, P. and Connelly, G. (1985). Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3):169–187.

Cahn, J. (1995). The effect of pitch accenting on pronoun referent resolution. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 290–292.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Computer Graphics '94*, pages 413–420.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper and Row, New York.

Collier, G. (1985). *Emotional Expression*. Lawrence Erlbaum Associates.

Condon, W. and Osgton, W. (1967). A segmentation of behavior. *Journal of Psychiatric Research*, 5:221–235.

Condon, W. and Osgton, W. (1971). Speech and body motion synchrony of the speaker-hearer. In Horton, D. and Jenkins, J., editors, *The Perception of Language*, pages 150–184. Academic Press.

Condon, W. and Sander, L. (1974). Synchrony demonstrated between movements of the neonate and adult speech. *Child Development*, 45:456–462.

Cooper, W. E. and Paccia-Cooper, J. (1980). *Syntax and Speech*. Harvard University Press.

Couper-Kuhlen, E. (1984). A new look at contrastive intonation. In Watts, R. and Weidmann, U., editors, *Modes of Interpretation: Essays Presented to Ernst Leisi*, pages 137–158. Gunter Narr Verlag.

Cruttendon, A. (1986). *Intonation*. Cambridge University Press.

Culicover, P. and Rochemont, M. (1983). Stress and focus in English. *Language*, 59:123–165.

Dale, R. and Haddock, N. (1991). Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4):252–265.

Davis, J. and Hirschberg, J. (1988). Assigning intonational features in synthesized spoken discourse. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 187–193, Buffalo.

Dik, S. (1980). On the topology of focus phenomena. *GLOT Leids Taalkundig Bulletin*, 3(3,4):41–74. Referenced in Gussenhoven (1983).

Dittmann, A. and Llewellyn, L. (1967). The phonemic clause as a unit of speech decoding. *Journal of Personality and Social Psychology*, 6(3):341–349.

Dowty, D. (1988). Type raising, functional composition, and non-constituent conjunction. In Oehrle, R., Bach, E., and Wheeler, D., editors, *Categorial Grammars and Natural Language Structures*, pages 153–198. Reidel, Dordrecht.

Duncan, S. (1973). Toward a grammar for dyadic conversation. *Semiotica*, 9:29–46.

Duncan, S. (1974a). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3:161–180.

Duncan, S. (1974b). Some signals and rules for taking speaking turns in conversations. In Weitz, editor, *Nonverbal Communication*. Oxford University Press.

Ekman, P. (1989). The argument and evidence about universals in facial expressions of emotion. In Wagner, H. and Manstead, A., editors, *Handbook of Social Psychophysiology*, pages 143–164. Wiley, Chichester; New-York.

Elhadad, M. (1993). *Using Argumentation to Control Lexical Choice: A Functional Unification Implementation*. PhD thesis, Columbia University.

Elhadad, M., McKeown, K., and Robin, J. (1996). Floating constraints in lexical choice. *Computational Linguistics*. To appear.

Elhadad, M. and Robin, J. (1992). Controlling content realization with functional unification grammars. In Dale, R., Hovy, E., Rosner, D., and Stock, O., editors, *Aspects of Automated Natural Language Generation*, pages 89–104. Springer Verlag, Berlin.

Engdahl, E. and Vallduví, E. (1994). Information packaging and grammar architecture: A constraint-based approach. In Engdahl, E., editor, *Integrating Information Structure into Constraint-Based and Categorial Approaches (DYANA-2 Report R.1.3.B)*. CLLI, Amsterdam.

Essa, I. A. (1994). *Analysis, Interpretation, and Synthesis of Facial Expressions*. PhD thesis, MIT, Media Laboratory, Cambridge, MA.

Fuchs, A. (1984). Deaccenting and default accent. In Gibbon, D. and Richter, H., editors, *Intonation, Accent and Rhythm: Studies in Discourse Phonology*, pages 134–164. De Gruyter, Berlin.

Gatewood, J. and Rosenwein, R. (1981). Interactional synchrony: Genuine or spurious? a critique of recent research. *Journal of Nonverbal Behavior*, 6(1):12–27.

Gerdeman, D. and Hinrichs, E. (1990). Functor–driven natural language generation with categorial unification grammars. In *COLING 90: Proceedings of the 13th International Conference on Computational Linguistics*, pages 145–150, Helsinki.

Goldsmith, J. (1976). *Autosegmental Phonology*. PhD thesis, Massachusetts Institute of Technology.

Grant, E. (1969). Human facial expression. *Man*, 4:525–536.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1986). Towards a computational theory of discourse interpretation. Unpublished manuscript.

Grosz, B. J. and Sidner, C. L. (1986). Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204.

Gussenhoven, C. (1983a). Focus, mode and the nucleus. *Journal of Linguistics*, 19:377–417.

Gussenhoven, C. (1983b). *On the Grammar and Semantics of Sentence Accent*. Foris, Dodrecht.

Hadar, U., Steiner, T., Grant, E., and Rose, F. C. (1983a). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(2):117–129.

Hadar, U., Steiner, T., Grant, E., and Rose, F. C. (1983b). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46.

Hadar, U., Steiner, T., and Rose, F. (1984). The relationship between head movements and speech dysfluencies. *Language and Speech*, 27(4):333–342.

Hajičová, E. (1987). Focussing: a meeting point of linguistics and artificial intelligence. In Jorrand, P. and Sgurev, V., editors, *Artificial Intelligence II: Methodology, Systems, Applications*, pages 311–321. Elsevier Science Publishers, North Holland.

Hajičová, E. and Sgall, P. (1987). The ordering principle. *Journal of Pragmatics*, 11:435–454.

Hajičová, E. and Sgall, P. (1988). Topic and focus of a sentence and the patterning of a text. In Petofi, J., editor, *Text and Discourse Constitution*. De Gruyter, Berlin.

Halliday, M. (1967). *Intonation and Grammar in British English*. Mouton, The Hague.

Halliday, M. (1970). Language structure and language function. In Lyons, J., editor, *New Horizons in Linguistics*, pages 140–165. Penguin.

Hardt, D. (1993). *Verb Phrase Ellipsis: Form, Meaning and Processing*. PhD thesis, University of Pennsylvania.

Heim, I. (1983). File change semantics and the familiarity theory of definiteness. In Bauerle, R., Schwarze, C., and von Stechow, A., editors, *Meaning, Use and Interpretation of Language*, pages 164–189. W. de Gruyter, Berlin.

Hirschberg, J. (1990). Accent and discourse context: Assigning pitch accent in synthetic speech. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 952–957.

Hoffman, B. (1995). *The Computational Analysis of the Syntax and Interpretation of 'Free' Word Order in Turkish*. PhD thesis, University of Pennsylvania, Philadelphia.

Houghton, G. (1986). *The Production of Language in Dialogue: a Computational Model*. PhD thesis, University of Sussex.

Houghton, G. and Isard, S. (1987). Why to speak, what to say and how to say it. In Morris, P., editor, *Modelling Cognition*. Wiley.

Houghton, G. and Pearson, M. (1988). The production of spoken dialogue. In Zock, M. and Sabah, G., editors, *Advances in Natural Language Generation: An Interdisciplinary Perspective, Vol. 1*. Pinter Publishers, London.

Hovy, E. (1988). Planning coherent multisentential text. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 163–169, Buffalo.

Hovy, E. (1992). Natural language generation. Technical Report RL-TR-92-273, Rome Laboratory Air Force Materiel Command.

Hovy, E. (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63:341–385.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA.

Jacobs, J. (1991). On the semantics of modal particles. In Abraham, W., editor, *Discourse particles : descriptive and theoretical investigations on the logical, syntactic, and pragmatic properties of discourse particles in German*. John Benjamins, Amsterdam.

Joshi, A. K. (1985). How much context-sensitivity is required to provide reasonable structural descriptions: Tree-adjoining grammars. In Dowty, D., Karttunen, L., and Zwicky, A., editors, *Natural Language Parsing: Psycholinguistic, Computational and Theoretical Perspectives*, pages 206–350. Cambridge University Press, New York.

Joshi, A. K. and Weinstein, S. (1981). Control of inference: Role of some aspects of discourse structure – centering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 385–387, Vancouver.

Kalra, P., Mangili, A., Magnenat-Thalmann, N., and Thalmann, D. (1991). SMILE: A multilayered facial animation system. In Kunii, T., editor, *Modeling in Computer Graphics*. Springer-Verlag.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

Kendon, A. (1972). Some relationships between body motion and speech. In Siegman, A. and Pope, B., editors, *Studies in Dyadic Communication*, pages 177–210. Pergamon Press, New-York.

Kendon, A. (1974). Movement coordination in social interaction: Some examples described. In Weitz, editor, *Nonverbal Communication*. Oxford University Press.

Krifka, M. (1992). A compositional semantics for multiple focus. In Jacobs, J., editor, *Informationsstruktur and Grammatik*, pages 17–53. Westdeutcher Verlag.

Kuno, S. (1976). Subject, theme and speaker's empathy: A reexamination of relativization phenomena. In Li, C., editor, *Subject and Topic*, pages 417–444. Academic Press, New York.

Ladd, D. R. (1980). *The Structure of Intonational Meaning*. Indiana University Press, Bloomington.

Ladd, D. R. (1983). *Even*, focus and normal stress. *Journal of Semantics*, 2:157–70.

Lakoff, G. (1971). Presupposition and relative well-formedness. In Steinberg and Jakobovits, editors, *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*, pages 329–340. Cambridge University Press.

Lakoff, G. (1972). The global structure of the nuclear stress rule. *Language*, 48:285–303.

Liberman, M. (1975). *The Intonational System of English*. PhD thesis, Massachusetts Institute of Technology.

Liberman, M. and Buchsbaum, A. L. (1985). Structure and usage of current Bell Labs text to speech programs. Technical Memorandum TM 11225-850731-11, AT&T Bell Laboratories.

Liberman, M. and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In Aranoff, M. and Oehrle, R., editors, *Language Sound Structure: Studies in Phonology Presented to Morris Halle*. MIT Press, Cambridge, MA.

Lyons, J. (1977). *Semantics, Volume II*. Cambridge University Press.

Mann, W. and Thompson, S. (1986). Rhetorical structure theory: Description and construction of text structures. In Kempen, G., editor, *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, pages 279–300. Kluwer Academic Publishers, Boston.

McDonald, D. D. (1986). Description directed control: Its implications for natural language generation. In Grosz, B., Jones, K. S., and Webber, B., editors, *Readings in Natural Language Processing*, pages 519–538. Morgan Kaufmann.

McKeown, K., Kukich, K., and Shaw, J. (1994). Practical issues in automatic documentation generation. In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*, pages 7–14, Stuttgart. Association for Computational Linguistics.

McKeown, K. R. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, Cambridge.

McKeown, K. R. (1986). Discourse strategies for generating natural language text. In Grosz, B., Jones, K. S., and Webber, B., editors, *Readings in Natural Language Processing*, pages 479–500. Morgan Kaufmann.

Meteer, M. (1991). Bridging the generation gap between text planning and linguistic realization. *Computational Intelligence*, 7(4):296–304.

Monaghan, A. (1991). *Intonation in a Text-to-Speech Conversion System*. PhD thesis, University of Edinburgh.

Moortgat, M. (1989). *Categorial Investigations*. Foris, Dodrecht.

Negroponte, N. (1995). *Being Digital*. Alfred A. Knopf, Inc., New York.

Nespor, M. and Vogel, I. (1989). *Prosodic Phonology*. Foris, Dordrecht.

Newman, S. (1946). On the stress system of English. *Word*, 2:171–187.

O'Connell, D. and Slaymaker, F. (1984). Evidence for the phonemic clause as an encoding unit. *Language and Communication*, 4(4).

Oehrle, R. (1988). Mulit-dimensional compositional functions as a basis for grammatical analysis. In Oehrle, R., Bach, E., and Wheeler, D., editors, *Categorial Grammars and Natural Language Structures*, pages 349–390. Reidel, Dordrecht.

O'Shaughnessy, D. (1977). Fundamental frequency by rule for a text-to-speech system. In *Proceedings of the Internationsl Conference on Acoustics, Speech and Signal Processing*, pages 568–570, New York. IEEE Cat. No. 77CH1197-3 ASSP.

Paris, C. (1987). *The Use of Explicit Models in Text Generation: Tailoring to a User's Level of Expertise*. PhD thesis, Columbia University.

Pelachaud, C., Badler, N., and Steedman, M. (1995). Generating facial expressions for speech. *Cognitive Science*. to appear.

Pelachaud, C. and Prevost, S. (1994). Sight and sound: Generating facial expressions and spoken intonation from context. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 216–219, New Paltz, NY.

Pelachaud, C. and Prevost, S. (1995a). Coordinating vocal and visual parameters for 3D virtual agents. In *Proceedings of the 2nd Eurographics Workshop on Virtual Environments*, Monte Carlo.

Pelachaud, C. and Prevost, S. (1995b). Talking heads: Physical, linguistic and cognitive issues in facial animation. Course Notes for Computer Graphics International '95.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology. Distributed by Indiana University Linguistics Club, Bloomington, IN.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In Cohen, P., Morgan, J., and Pollock, M., editors, *Intentions in Communication*, pages 271–312. MIT Press, Cambridge, MA.

Pitrelli, J., Beckman, M., and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Yokohama.

Polanyi, L. (1988). A formal model of the structure of discourse. In *Journal of Pragmatics*, pages 601–638.

Postal, P. (1964). Limitations of phrase structure grammars. In Fodor, J. and Katz, J., editors, *The Structure of Language*, pages 137–151. Prentice Hall, Englewood Cliffs, NJ.

Postal, P. (1972). Some further limitations of interpretive theories of anaphora. *Linguistic Inquiry*, 3:349–371.

Prevost, S. (1995). Contextual aspects of prosody in monologue generation. In *IJCAI Workshop on Context in Natural Language Processing.*

Prevost, S. and Steedman, M. (1993a). Generating contextually appropriate intonation. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, pages 332–340, Utrecht.

Prevost, S. and Steedman, M. (1993b). Using context to specify intonation in speech synthesis. In *Proc. 3rd European Conf. Speech Communication and Technology (EUROSPEECH)*, pages 2103–2106, Berlin.

Prevost, S. and Steedman, M. (1994a). Information based intonation synthesis. In *Proceedings of the Human Language Technology Workshop*, pages 193–198.

Prevost, S. and Steedman, M. (1994b). Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153.

Prince, E. F. (1981). Towards a taxonomy of the given/new distinction. In Cole, P., editor, *Radical Pragmatics*, pages 223–255. Academic Press, London.

Prince, E. F. (1986). On the syntactic marking of the presupposed open proposition. *Journal of the Chicago Linguistic Society*, pages 208–222.

Rambow, O. and Korelsky, T. (1992). Applied text generation. In *Proceedings of the Third Conference on Applied Natural Language Processing (ANLP-1992)*, pages 40–47.

Reinhart, T. (1981). Pragmatics and linguistics: an analysis of sentence topic. *Philosophica*, 27:53–94.

Reiter, E. (1991). A new model of lexical choice for nouns. *Computational Intelligence*, 7(4):240–251.

Reiter, E. (1994). Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 163–170.

Reiter, E. and Dale, R. (1992). A fast algorithm for the generation of referring expressions. In *COLING 92: Proceedings of the 14th International Conference on Computational Linguistics*, pages 232–238.

Reiter, E. and Mellish, C. (1992). Using classification to generate text. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 265–272.

Robin, J. (1993). A revision-based generation architecture for reporting facts in their historical context. In Horacek, H. and Zock, M., editors, *New Concepts in Natural Language Generation: Planning, Realization and Systems*, pages 238–265. Pinter Publishers, New York.

Robin, J. (1994). *Revision-Based Generation of Natural Language Summaries Providing Historical Background: Corpus-Based Analysis, Design, Implementation and Evaluation.* PhD thesis, Columbia University.

Robin, J. and McKeown, K. (1993). Corpus analysis for revision-based generation of complex sentences. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 365–373.

Rochemont, M. (1986). *Focus in Generative Grammar.* John Benjamins, Philadelphia.

Rooth, M. (1985). *Association with Focus.* PhD thesis, University of Massachusetts, Amherst.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1:75–116.

Sagisaka, Y. (1990). On the prediction of global F0 shape for Japanese text-to-speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 325–328.

Scherer, K. R. (1980). The functions of nonverbal signs in conversation. In St. Clair, R. and Giles, H., editors, *The Social and Physiological Contexts of Language*, pages 225–243. Lawrence Erlbaum Associates.

Schmerling, S. F. (1976). *Aspects of English Sentence Stress*. University of Texas Press, Austin.

Selkirk, E. (1984). *Phonology and Syntax*. MIT Press, Cambridge, MA.

Shieber, S. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343.

Shieber, S. and Schabes, Y. (1991). Generation and synchronous tree-adjoining grammars. *Computational Intelligence*, 4:220–228.

Sibun, P. (1991). *The Local Organisation and Incremental Generation of Text*. PhD thesis, University of Massachusetts.

Sibun, P. (1992). Generating text without trees. *Computational Intelligence: Special Issue on Natural Language Generation*, 8(1).

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling english prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 867–870, Banff.

Smadja, F. and McKeown, K. (1991). Using collocations for language generation. *Computational Intelligence*, 7(4):229–239.

Sperber, D. and Wilson, D. (1986). *Relevance: Communication and Cognition*. Harvard University Press, Cambridge, MA.

Steedman, M. (1985). Dependency and coordination in the grammar of Dutch and English. *Language*, 61:523–568.

Steedman, M. (1990a). Gapping as constituent coordination. *Linguistics and Philosophy*, 13:207–263.

Steedman, M. (1990b). Structure and intonation in spoken language understanding. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 9–17, Pittsburgh.

Steedman, M. (1991a). Structure and intonation. *Language*, pages 260–296.

Steedman, M. (1991b). Surface structure, intonation, and focus. In Klein, E. and Veltmann, F., editors, *Natural Language and Speech: Proceedings of the Symposium, ESPRIT Conference*, pages 21–38, Brussels.

Steedman, M. (1991c). Surface structure, intonation and meaning in spoken language. In Bates, M. and Weischedel, R., editors, *Challenges in Natural Language Processing*. Cambridge University Press.

Steedman, M. (1991d). Type-raising and directionality in categorial grammar. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 71–78, Berkeley.

't Hart, J. and Cohen, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics*, 1:309–327.

't Hart, J. and Collier, R. (1975). Integrating different levels of phonetic analysis. *Journal of Phonetics*, 3:235–255.

Terken, J. (1984). The distribution of accents in instructions as a function of discourse structure. *Language and Structure*, 27:269–289.

Terken, J. and Hirschberg, J. (1994). Deaccentuation of words representing 'given' information: Effects of persistence of grammatical function and surface position. *Language and Speech*, 37(2):125–145.

Ullman, S. (1984). Visual routines. In Pinker, S., editor, *Visual Cognition*, pages 97–161. Elsevier Science Publishers, Amsterdam.

Vallduví, E. (1990). *The Informational Component*. PhD thesis, University of Pennsylvania, Philadelphia.

Vilkuna, M. (1989). *Free Word Order in Finnish: Its Syntax and Discourse Functions*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Walker, M., Iida, M., and Cote, S. (1990). Centering in Japanese discourse. In *COLING 90: Proceedings of the 13th International Conference on Computational Linguistics*, pages 1–6, Helsinki.

Walker, M., Iida, M., and Cote, S. (1994). Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232.

Walker, M. and Trimboli, C. (1983). The expressive function of the eye flash. *Journal of Nonverbal Behavior*, 8(1):3–13.

Wang, M. Q. and Hirschberg, J. (1991). Predicting intonational phrasing from text. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 285–292, Berkeley.

Ward, G. and Hirschberg, J. (1985). Implicating uncertainty: the pragmatics of fall-rise intonation. *Language*, pages 747–776.

Webber, B., Rymon, R., and Clarke, J. (1992). Flexible support for trauma management through goal-directed reasoning and planning. *Artificial Intelligence in Medicine*, 4(2):145–163.

Webbink, P. (1986). *The Power of the Eyes*. Springer Publishing Company.

Weir, D. (1988). *Characterizing Mildly Context-Sensitive Grammar Formalisms*. PhD thesis, University of Pennsylvania.

Wiemann, J. and Knapp, M. (1975). Turn-taking in conversations. *Journal of Communication*, pages 75–92.

Yacoob, Y. and Davis, L. (1994). Computing spatio-temporal representations of human faces. In *Computer Vision and Pattern Recognition Conference*, pages 70–75. IEEE Computer Society.

Young, S. and Fallside, F. (1979). Speech synthesis from concept: a method for speech output from information systems. *Journal of the Acoustical Society of America*, 66:695–695.

Zacharski, R., Monaghan, A., Ladd, D., and Delin, J. (1993). BRIDGE: Basic research on intonation in dialogue generation. Technical report, HCRC: University of Edinburgh. Unpublished manuscript.