NAIST-IS-DT9661010

Doctor's Thesis

A Corpus-based Study on Conversational Interaction in Japanese: Discourse Structures, Turn-Taking, and Backchannels

Hanae Koiso

July 27, 1998

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

Doctor's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
DOCTOR of SCIENCE

Hanae Koiso

Thesis committee: Yuji Matsumoto, Professor
Kiyohiro Shikano, Professor
Wilhelm Nicholas Campbell, Professor
Yasuharu Den, Associate Professor

A Corpus-based Study on Conversational Interaction in Japanese: Discourse Structures, Turn-Taking, and Backchannels*

Hanae Koiso

Abstract

In conversation, conversants interact with each other regularly and smoothly, although conversational organization, such as changes of speakers and topics, is not fixed in advance. What regulates and facilitates such smooth interaction? In the last few decades, the mechanisms underlying such smooth interaction have become a central research concern. Researchers in various fields, such as discourse analysis, sociolinguistics, conversation analysis, social psychology, philosophy, pragmatics and anthropology, observed real conversational interactions very closely, becoming recognize the importance of linguistic, physical, and social contexts in which interactions occur. Such conversational contexts, provided by the conversants, potentially carry information about conversational organization, namely, meta-level messages, which in turn regulate interaction among the conversants.

In this dissertation, we discuss the following topics on conversational interaction in Japanese:

1. The elucidation of the function of syntactic and prosodic features as contextualization cues to conversational organization, including

*Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9661010, July 27, 1998.

- Global structural elements: discourse structures, and
- Local structural elements: turn-taking and backchannels

2. The construction of a model of a conversant's action concerning turn-taking

The study is based on analyses of a spontaneous dialogue corpus in Japanese, where linguistic and para-linguistic features are, automatically or semi-automatically, extracted from high-quality recorded speech materials and their correlations to the phenomena are statistically investigated.

In Chapter 3, we start with exploring the global structural elements, discourse structures. We investigate the relationship between discourse structures and changes in the speech rate in several conditions, elucidating the informational potentials of dynamic speech rate in dialogues. We claim that changes in the speech rate have definite potential in signaling the structure of information collaboratively constructed by the conversants of a conversation.

In Chapter 4, we pay attention to the local structural elements, turn-taking and backchannels. We explore the predictive powers of prosodic and syntactic features of the speaker's speech for discriminating turn-taking categories and for discriminating backchannel categories, elucidating the function of these features as contextualization cues to turn-taking and backchannels. We further discuss the interrelationship between prosody and syntax in the decision process of these phenomena.

In Chapter 5, following the descriptive studies presented in the previous two chapters, we develop a model of a conversant's action which tries to understand the mechanism underlying turn-taking phenomena, without supposing direct use of a signaling system. We show that our non-signaling model can account for smooth transitions of turn as well as the regularities observed in wider range of conversational interactions.

In Chapter 6, we summarize the dissertation and describe future directions.

Keywords:

discourse structures, turn-taking, backchannels, prosody, syntax, spontaneous dialogues

Acknowledgements

I am sincerely grateful to a lot of colleagues and friends who helped make this work possible.

First of all, I would like to express my gratitude to the members of my committee, Professor Yuji Matsumoto, Professor Kiyohiro Shikano, Professor Nick Campbell, and Associate Professor Yasuharu Den. Prof. Matsumoto provided me the opportunity to conduct this research, and gave me valuable suggestions and advice as well as continuous encouragement. Prof. Shikano and Prof. Campbell gave me valuable suggestions and helpful comments. Prof. Den has been always generous with time and advice, and had a great influence on my work. I benefited a lot from his insightful guidance and inspiration as well as his encouragement and patience.

I am also indebted to Dr. Ryohei Nakatsu and Dr. Yasuhiro Katagiri, who provided the opportunity for me to work at ATR Media Integration & Communications Research Laboratories. Stimulating discussions with Dr. Katagiri and with Prof. Atsushi Shimojima, who was working with me at ATR until this spring, has allowed me to get much further than it often seemed I could.

I would like thank to Prof. Syun Tutiya, who was my supervisor when I was at the laboratory of Cognitive and Information Sciences, Chiba University. He first provided the opportunity for me to study speech communication through the Japanese Map Task Corpus project, and gave me a lot of valuable advice and suggestion as well as constructive criticisms during many lively discussions. I also wish to thank Prof. Akira Ichikawa and Dr. Yasuo Horiuchi of Chiba University for sharing their time and their wisdom.

I am grateful to many members of Prof. Matsumoto's laboratory and of ATR.

My thanks are especially to Dr. Takehito Utsuro, Dr. Takashi Miyata, Dr. Michio Okada, Hiroaki Noguchi, Noriko Suzuki, Yugo Takeuchi, and Yoshinori Sakane for fruitful discussions and technical supports, and to Dr. Kristiina Jokinen, Dr. Patrick Healey, and Dr. Marfu Hasan for proofreading and helpful comments and suggestions.

Many other people deserve acknowledgement on the academic and technical front. Among them, I am especially grateful to Prof. Masato Ishizaki and Dr. Marc Swerts for their stimulating discussions and helpful advice, and to Miwako Kurihara, Izumi Ito, Kazuhisa Kiriyama, Yasuko Fukuda, and the members of the Japanese Map Task Corpus project for their technical help.

Personal thanks are due to Yuko Den and Koichiro Hajiri for their continuous encouragements, supports, patience, and fortitude over the past few years.

Finally, my heartfelt thanks go to my parents, Tadao and Sachie Koiso, for being so supportive and encouraging throughout my academic career.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In conversation, conversants interact with each other regularly and smoothly, although conversational organization, such as changes of speakers and topics, is not fixed in advance. What regulates and facilitates such smooth interaction? In the last few decades, the mechanisms underlying such smooth interaction have become a central research concern. Researchers in various fields, such as discourse analysis, sociolinguistics, conversation analysis, social psychology, philosophy, pragmatics and anthropology, observed real conversational interactions very closely, recognizing the importance of linguistic, physical, and social *contexts* in which interactions occur. Gumperz and Cook-Gumperz (1982) states that, apart from syntactic grammar, there is another level of information which must be shared by conversants if communication is to take place; this sort of information can be carried through the contexts.

People engaged in a conversation not only express, by means of uttering words and sentences, what they intend to get across to other conversants about the topics of discourse, but also express, both intentionally and unintentionally, messages concerned with the context in which the conversation is taking place. This second type of expression in many ways directs how the first type of expression is to be interpreted by the conversants. The second type of expression is usually carried out through non-referential and non-propositional means such as prosody,

gesture/posture, gaze, and expressions peculiar to dialogues. For example, various events, such as the direction in which the conversants gaze, the distance between the conversants, and how often the hearer utters backchanneling signals, can carry various information of this type. Thus, we exchange *meta-level* messages at any time, and exchanges of these messages could be thought as one of the factors making complicated conversational interactions more smooth. Gumperz and other sociolinguists (Gumperz, 1989, 1991; Erickson & Shultz, 1981; Auer & Luzio, 1992) introduced the notion of *contextualization* to capture these processes in everyday language use, and called the signals involved in expressions of the second type *contextualization cues.*

Thus, conversational contexts, provided by the conversants, carry information about conversational organization, namely, meta-level messages, which in turn regulate interaction among the conversants. The exploration of the relation between contexts and conversational organization, therefore, would provide a foundation for studies on the mechanisms underlying smooth interaction. In this dissertation, based on analyses of a spontaneous dialogue corpus in Japanese, we investigate the relationship between several phenomena concerning conversational organization and linguistic and para-linguistic contexts. This will not only make clear the signaling potentials of these linguistic and para-linguistic contextualization cues to the phenomena, but also help us to understand how conversants are acting in a conversation to achieve smooth interaction with access to those contexts. Furthermore, it helps to develop advanced spoken dialogue systems on computers, because present dialogue systems completely lack this very aspect of conversational interaction, that is, exchanges of meta-level messages based on linguistic, physical, and social contexts.

We take up three phenomena concerning conversational organization, and investigate the signaling potentials of several prosodic and syntactic features of the speaker's speech with respect to them. First, we focus on global structural elements, *discourse structures.* As Maynard (1989) mentioned, based on the global structure of conversational interaction, the conversants can understand the current status of interaction from a holistic perspective, making it possible for the conversants to predict some sorts of future events like when the next topic starts. Prosodic and syntactic features of the speaker's speech are thought to be

of help in understanding discourse structures; these features could carry various information such as when the current topic terminates and the relation of the current topic to the previous one. Next, we focus on local structural elements, *turn-taking* and *backchannels*. In conversations, although who speaks and when are not predetermined, speaking turns usually change regularly and smoothly, and backchannels are produced by the hearer on appropriate occasions. We can, therefore, assume a system of signals which regulate the process of turn-taking and backchannels. Our investigation into the relation of these phenomena to prosodic and syntactic features of the speaker's speech makes one aspect of conversational contexts more clear.

1.2 Previous Works

In this section, we give an overview of the previous works on discourse structures, turn-taking, and backchannels.

1.2.1 Studies on Discourse Structures

It has been supposed that a discourse can be partitioned into segments, and each segment can embed and be embedded in other segment, resulting in a hierarchical structure (Grosz, Pollack, & Sidner, 1989). Thus, the structure of a discourse can be very complicated, and therefore, conversants are supposed to have various clues for understanding the structure of a discourse. Among them, linguistic clues have been mainly investigated in computational linguistics, and prosodic ones in speech engineering. In this section, we give an overview of studies on discourse structures and their relation to linguistic and prosodic clues.

1.2.1.1 Discourse Structures

Interactive dialogues manifest their structures at various levels: local coordinative structures (e.g., adjacency pairs (Schegloff & Sacks, 1973)), dialogue topic structures, and higher level intentional structures on both discourse- and task-related plans. These structures form a hierarchy of information structures for

dialogues. Some researchers consider a dialogue as a rule-governed game, and divide the dialogue into "moves" according to the types of their contributions to the game process (Coulthard, 1977; Stenström, 1994; Carletta et al., 1997). Some researchers provide a foundation for discourse structures based on the intention of the speaker, and divide the dialogue into "discourse segments," each of which consists of three types of components, the intentional, attentional, and linguistic components (Grosz & Sidner, 1986). Other researchers focus on a particular goal of a dialogue, and divide the dialogue into segments according to their contributions to this goal. In particular, Clark and Schaefer (1989) and Traum (1994) are interested in the sharing by the speaker and the hearer, or *grounding*, of information presented by the speaker as a goal for a dialogue. They segmented the dialogue from this informational point of view.

1.2.1.2 Discourse Structures and Linguistic Features

Since, in computational linguistics, discourse structures are supposed to play an important role in the understanding of natural language by computers and by humans, the relationship between discourse structures and various linguistic features has been extensively investigated. In particular, *cue phrases* or *clue words*, such as "now," "anyway," and "that is" in English, were widely studied (Cohen, 1984; Mann & Thompson, 1986; Grosz & Sidner, 1986; Litman & Allen, 1987; Hirschberg & Litman, 1993). These phrases or words are suggested to indicate the boundaries of discourse structures or relationships between discourse segments (Grosz & Sidner, 1986).

Let us see the following example (Hirschberg & Litman, 1993, p. 504).

> If the system attempts to hold rules, *say* as an expert database for an expert system, *then* we expect it not only to hold the rules but to in fact apply them for us in appropriate situations.

Hirschberg and Litman (1993) pointed out that "say" indicates the beginning of a subtopic, and "then" a return from that subtopic to the major topic.

As Grosz et al. (1989) mentioned, these cue phrases do "not make a direct semantic contribution to an utterance, but instead convey information about the

structure of the discourse containing the utterance" (Grosz et al., 1989, p. 443). In this respect, we can regard cue phrases as contextualization cues to the structure of a discourse.

1.2.1.3 Discourse Structures and Prosodic Features

Speech researchers have examined the relation of discourse structures to the following prosodic features based on analyses of spontaneous monologues and dialogues.

Pitch Range It has been reported by many researchers that the pitch range of the speaker is expanded at the beginning of a new topic (Butterworth, 1974; Brown, Currie, & Kenworthy, 1980; Hirschberg & Pierrehumbert, 1986; Silverman, 1987). For example, Brown et al. (1980) observed that at the beginning of topics, subjects tend to speak relatively high in their pitch range, while they tend to compress their range at the end. Hirschberg and Pierrehumbert (1986) also reported that a larger amount of lowering at the end of an utterance brings a stronger sense of completion to the current topic.

Pitch Value Absolute pitch values have also been reported to have a relation to discourse structures. Hirschberg and Nakatani (1996) analyzed read and spontaneous monologues, reporting that, in each type of monologue, pitch values tend to be higher at the beginning parts of discourse segments than at the intermediate and final parts. In Japanese, Nakajima and Tsukada (1997) reported that pitch value is higher at the beginning of topics, and that when the current utterance is subordinative to the previous or following utterance, the pitch value of the current utterance is low. From this result, they claimed that the speaker uses higher pitch to indicate to the hearer a shift to a new topic, while lower pitch indicates elaboration of the current topic.

Intonation It has been pointed out that there is correspondence between discourse structures and intonation or boundary tones (Hirschberg & Pierrehumbert, 1986; Pierrehumbert & Hirschberg, 1990; Swerts & Geluykens, 1994). Pierrehumbert and Hirschberg (1990) suggested that high boundary tones are used by the

speaker to signal that the current utterance should be interpreted with respect to subsequent utterances, meaning that they do not felicitously end discourse segments; while low boundary tones convey the opposite. Swerts and Geluykens (1994) also found that high boundary tones are associated with the non-finality of topical units, and low boundary tones with finality, insisting that low contours signal that a topical unit has been completed, while high contours signal that there is still more to come on the same topic.

Amplitude Several researchers reported that there is some relationship between amplitude and discourse structures (Brown et al., 1980; Hirschberg & Nakatani, 1996). Brown et al. (1980) found that amplitude increases at the beginning of topics and decreases at the end. Hirschberg and Nakatani (1996) also observed that amplitude tends to decrease toward the end of a discourse segment.

Pausal Duration It has also been reported by many researchers that pauses preceding topic shifts tend to be longer than topic-internal pauses (Lehiste, 1979; Brown et al., 1980; Swerts & Geluykens, 1994; Hirschberg & Nakatani, 1996; Den & Koiso, 1998). For example, Swerts and Geluykens (1994) found that, in Dutch monologues, pauses occur at all transitions between topical units, and that pauses between topical units are longer than in other locations. Den and Koiso (1998), analyzing task-oriented dialogues and casual conversations in Japanese, found that, in each style of conversation, pauses at the boundaries of game units are longer than those within game units.

Speech Rate Speech rate has also been pointed out to have strong relation to discourse structures. Brubaker (1972) studied average speech rates in individual sentences in read monographs, and found that speech tends to accelerate as it approaches the end of a paragraph. He also showed that speech gradually accelerates within individual sentences. Koopmans-van Beinum and van Donzel (1996) studied speech rates in inter-pausal units and their relations to paragraph boundaries in read monographs, and found that the initial segment of speech after a paragraph boundary is slower than the intermediate segment of a paragraph. Hirschberg and Nakatani (1996) claimed that average speech rates over individ-

ual intermediate phrases tend to be greater toward the end of discourse segments both in read and spontaneous monologues, while speech tends to be slower in the middle parts of discourse structures in spontaneous monologues. They also studied speech rate changes over adjacent phrases, and found that speech tends to decelerate at the beginning of discourse segments.

Thus, such prosodic features as pitch range, pitch values, intonation, amplitude, pausal duration, and speech rate, as well as linguistic features, are supposed to function as contextualization cues to discourse structures.

1.2.2 Studies on Turn-Taking

In conversation, conversants take speaking turns and respond with backchannels regularly and smoothly, although who speaks and when are not decided in advance. To understand the mechanisms underlying these phenomena, numerous attempts have been made in various fields such as conversation analysis (Sacks, Schegloff, & Jefferson, 1974; Goodwin, 1981; Schegloff, 1982), social psychology (Kendon, 1967; Duncan & Fiske, 1977; Beattie, 1983), and discourse analysis (Yngve, 1970; Maynard, 1989; Ford & Thompson, 1996). Among these, the signal-based approach presented by Duncan and Fiske (1977) and the rule-based approach proposed by Sacks et al. (1974) are two influential theories. In this section, we give an overview of these two approaches as well as studies on the context of turn-taking.

1.2.2.1 Signal-Based Approach

Duncan (1972) and Duncan and Fiske (1977) proposed a turn-taking mechanism in which conversants regulate their turns by exchanging various signals. They suppose these signals to indicate each conversant's *state* with regard to the speaking turn (Duncan, 1972, p. 285).

As basic signals, Duncan (1972) reported (a) turn-yielding signals by the speaker, (b) attempt-suppressing signals by the speaker, and (c) backchannel signals by the hearer. Signals, which are composed of various *cues* such as intonation and body motion, work under a system of rules governing their interpretation.

(a) Turn-yielding signals

Rule When the speaker gives this type of signal, the hearer may take her speaking turn. If the hearer acts to take her turn in response to the signal, the speaker immediately yields his turn.

Cue A turn-yielding signal consists of at least one of a set of six cues: (1) rising or falling intonation patterns, (2) lengthening of the final syllable or on the stressed syllable of a terminal clause, (3) the termination of any hand gesticulation, (4) the appearance of one of several stereotyped expressions (e.g., "but uh," "or something," and "you know"), (5) a drop in pitch and/or loudness with one of the stereotyped expressions, and (6) the completion of a grammatical clause.

(b) Attempt-suppressing signals

Rule When the speaker gives this type of signal, the hearer should not take her turn and the speaker continues his turn, regardless of the number of yielding cues concurrently being displayed.

Cue An attempt-suppressing signal consists of the gesticulations of the speaker's hands.

(c) Backchannel signals

Rule When the speaker is displaying a turn-yielding signal, a backchannel is often used by the hearer to avoid taking her turn. As a result, the speaker continues his turn.

Cue A backchannel signal consists of various cues such as (1) expressions like "uh huh" and "yeah," (2) head nods, (3) sentence completions, in which the hearer completed a sentence that the speaker had begun, (4) brief requests for clarification, and (5) restatement in a few words of an immediately preceding thought expressed by the speaker.

Duncan and Fiske (1977) statistically investigated the relationship between these cues and several states of turn transitions, and found strong correlations between them.

1.2.2.2 Rule-Based Approach

Sacks et al. (1974) proposed a turn-taking mechanism in which turns are regulated by a set of rules, providing for the allocation of a next turn to one conversant, and coordinating transfer so as to minimize gaps and overlaps. These rules are applied at *transition-relevance places*, which are possible completion points of basic units of turns. These basic units of turns are called *turn-constructional units*, including sentential, clausal, phrasal, and lexical constructions.

The following is a basic set of rules governing turn-taking (Sacks et al., 1974, p. 704):

1. For any turn, at the initial transition-relevance place of the initial turn-constructional unit:

   (a) If the current speaker selects a next speaker in the current turn,[1] then the current speaker must stop speaking, and the next speaker must speak next.

   (b) If the current speaker does not select a next speaker, then any other party may self-select, first speaker gaining rights to the next turn.

   (c) If the current speaker has not selected a next speaker, and no other party self-selects under option (b), then the current speaker may (but need not) continue.

2. If, at the initial transition-relevance place, neither rule (a) nor (b) has operated, and the current speaker has continued, then the rules (a)–(c) re-apply at the next transition-relevance place, and recursively at each next transition-relevance place, until transfer is effected.

Sacks et al. (1974) suggested that the turn-taking mechanism based on these rules could account for or is compatible with various phenomena occuring in conversation, such as (1) speaker-change recurs, (2) overwhelmingly, one conversant

[1]The most typical technique for selecting a next speaker is that the current speaker makes initiations, such as a question, an offer, a request, and a tagged assertion, with an address term directed to a particular conversant.

talks at a time, (3) occurrences of more than one conversant speaking simultaneously are common, but brief, and (4) turn order and turn size are not fixed, but vary.

1.2.2.3 The Context of Turn-Taking

One matter of importance in these two approaches is how the context of turn-taking is characterized by linguistic, para-linguistic, and/or non-linguistic features, namely, what kinds of turn-yielding cues exist, or what factors characterize the relevant places for a turn transition at the boundary of turn-constructional units.

Some researchers have proposed that changes of turn tend to occur at syntactic or grammatical completion points (Sacks et al., 1974; Duncan & Fiske, 1977; Orestrŏm, 1983; Ford & Thompson, 1996), while turns are likely to be held when grammatical units being constructed are not completed (Ball, 1975). Maynard (1989), who examined the relation of turn-taking to syntactic features in dyadic casual conversations in Japanese, suggested that changes of turn occur at grammatical completion points characterized by utterance final intonation contours and the following syntactic features (Maynard, 1989, pp. 145–146): (1) sentential units, including sentences with ellipses, (2) gerundive endings of verbs, (3) subordinate clause endings without corresponding main clauses, (4) postposed sentences, considered to be complete at the end of postposed elements, and (5) independent fillers, employed to fill a potential silence, when accompanied by a verb.

It has also been proposed that changes of turn are connected with some kinds of prosodic or para-linguistic features, such as rising or falling intonation at the end of clauses (Duncan & Fiske, 1977; Ford & Thompson, 1996; Hinds, 1978) and a rapid drop in loudness in conjunction with stereotyped expressions (Duncan & Fiske, 1977). For example, Inoue (1997) and Kôri (1997) proposed that rise-falling and flat-falling intonations are associated with continuations of the same speaker's turn, and Osaka (1988), developing his model of turn-taking based on Hidden Stochastic Petri Net, emphasized a role of energy.

Furthermore, some kinds of body motion have been pointed out to be related

to turn-taking. For example, Kendon (1967) reported that when the speaker's speech terminates with his gaze at the hearer, the hearer is more likely to respond without a pause than if it terminates without the speaker's gaze. Kendon (1967) and Duncan (1972) also reported that the speaker tends to terminate any hand gesticulation used during a speaking turn when he stops his speaking turns.

Thus, various syntactic, prosodic, and non-verbal features are supposed to have the function as contextualization cues to turn-taking.

1.2.3 Studies on Backchannels

Backchannels, or *aizuchi*, are short utterances, such as "yeah" and "uh huh" in English and "hai" and "ee" in Japanese, spoken by the hearer during the speaker's speech. They are reported to occur very frequently in Japanese conversations (Maynard, 1989). Thus, backchannels in Japanese seem to have strong influence on conversational interaction. Maynard (1989) pointed out several functions of backchannels in Japanese:

- Continuer,
- Display of understanding of content,
- Support toward the speaker's judgment,
- Agreement,
- Strong emotional response, and
- Minor addition, correction, or request for information.

In particular, the function of "continuer" is strongly related to turn-taking. This function was originally reported by Schegloff (1982); he argued that the hearer passes up opportunities for taking turns during moments in which the hearer could potentially take turns, and therefore a backchannel can be regarded as a *continuer* by which the hearer encourages the speaker to continue with the talk. The backchannel signals in the signal-based approach have also been treated in this line.

Several attempts have been made to characterize the context for the occurrence of backchannels. Erickson (1979) mentioned that there are moments where

the hearer is obliged to show more active listening responses, namely, some sorts of backchannels, while the speaker is speaking. He said that such "listening response relevant moments" seem to be signaled by cues in the speaker's speech. In order to characterize such moments, he investigated cues by the speaker such as syntactic juncture, intonation contour, changes in the tempo of speaking, gaze direction, and postural shifts.

For Japanese, Maynard (1989) reported that backchannels usually occur during the pause between clauses, which is sometimes marked by sentence-final particles and head movements. It has also been suggested that backchannels tend to appear in the context characterized by some kinds of prosodic features (Maynard, 1986; Mizutani, 1988; Imaishi, 1994; Ward, 1996). For example, in Japanese conversations, prosodic features such as rising, rise-falling, and flat-falling intonation contours serve as signals for inducing backchannels by the hearer (Imaishi, 1994).

Thus, some syntactic, prosodic, and non-verbal features are also supposed to have the function as contextualization cues to backchannels.

1.3 The Aim of the Dissertation

In the first part of the dissertation, in Chapters 3 and 4, we give descriptive accounts of conversational phenomena based on the idea of *signaling potentials* of exchanging meta-level messages. We investigate the relation of linguistic and para-linguistic contexts to conversational organization, including

- Global structural elements: discourse structures, and
- Local structural elements: turn-taking and backchannels.

The study is based on analyses of a spontaneous dialogue corpus in Japanese, where syntactic and prosodic features are, automatically or semi-automatically, extracted from high-quality recorded speech materials and their correlations to the phenomena are statistically investigated.

We elucidate the signaling potentials of these syntactic and prosodic cues to the global and local structural organizations. This part of the study is depicted in the upper part of Figure 1.1.

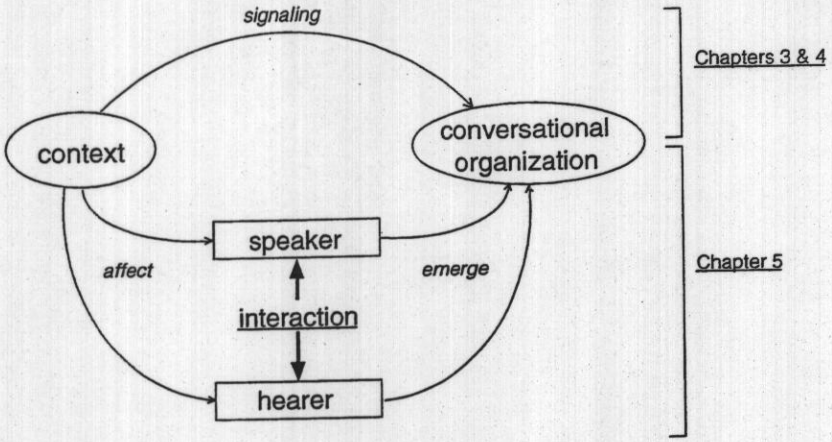


Figure 1.1: Overall design of the study.

In the latter part, in Chapter 5, we try to construct a model of a conversant's action to realize smooth interaction, particularly focusing on turn-taking. Although the analyses in Chapters 3 and 4 give us detailed descriptions of the phenomena, they do not make clear the actual processes taken by conversants. Thus, in Chapter 5, we explore the processes taken by conversants, providing a new model of turn-taking which accounts for the mechanism underlying smooth conversational interaction. The latter part of the study is depicted in the lower part of Figure 1.1.

1.3.1 Signaling Approach to Discourse Structures

Recently, the relation of discourse structures to various features has been examined. Several speech researchers have particularly focused on prosody, widely investigating the relation of discourse structures to prosody in *monologues*; while very few studies have so far been addressed to dialogues. It may be partly because of the difficulty of analyzing dialogues due to their interactive aspects. To fill this gap, we investigate the relationship between discourse structures and prosody in

dialogues, particularly focusing on local changes of speech rate.

It is widely held that, in monologues, speech tends to accelerate within discourse segments, while it tends to decelerate at the boundaries of segments (Brubaker, 1972; Koopmans-van Beinum & van Donzel, 1996; Hirschberg & Nakatani, 1996), that is, changes of speech rate could have potential in conveying information about discourse structures. Since dialogue has multiple speakers, bringing about various interactive aspects such as turn-taking and backchannels, some questions that have never been posed in monologue research arise; whether acceleration patterns are universal across single-speaker utterances and cross-speaker utterances; whether the insertion of backchannels by other parties change the speech rate characteristics within a single speaker's utterances. Thus, the relation between changes of speech rate and discourse structure in dialogues has posed various unresolved problems.

In Chapter 3, in order to extend our knowledge of interactive aspects of discourse structures, we investigate the relationship between discourse structures and dynamic speech rates in several conditions, elucidating the signaling potentials of varying speech rate in dialogues.

1.3.2 Signaling Approach to Turn-Taking and Backchannels

Several studies have investigated the relation of turn-taking and backchannels to syntactic and prosodic features in order to characterize the context provided by conversants. Little attention, however, has been given to the *interrelationships* among the features.

Researchers such as Oreström (1983) and Ford and Thompson (1996) investigated several features at once, such as syntactic, prosodic, physical, pragmatic and semantic features. They, however, merely analyzed the relationship between turn-taking and *each individual feature*, leaving the exploration of the interrelationships among the features untouched. Sacks et al. (1974) suggested that turn-constructional units are mainly characterized by syntax but are also affected by prosody. However, they did not discuss the interaction between syntax and prosody in detail. As Auer (1996) mentioned, the ways in which syntax

and prosody contribute to the construction of turns are very complex, and the interrelationships between syntax and prosody with respect to turn-taking and backchannels has left various unresolved problems.

In Chapter 4, we first explore the relation of turn-taking and backchannels to several syntactic and prosodic features of the speaker's speech, elucidating the function of these features as contextualization cues to turn-taking and backchannels. Then, we discuss the interaction between prosody and syntax in the decision process of turn-taking and backchannels.

1.3.3 Non-Signaling Approach to Turn-Taking

The analyses in Chapters 3 and 4 give us detailed descriptions of the phenomena. They, however, do not make clear the actual processes taken by conversants to achieve smooth interaction. In Chapter 5, we try to construct a model of a conversant's action, which accounts for the mechanism of such smooth interaction.

Two models can be hypothesized: one is the *signal-based*, or *code* model, and the other is the *non-signaling* model. In the signal-based model, it is assumed that the speaker encodes a meta-level message concerning conversational organization into the current context following a code system, namely, a system of signaling relations, and that the hearer decodes the message from the context following the same code system. In the non-signaling model, on the other hand, it is not assumed that conversants directly make use of such a system of signaling relations. Instead, it is assumed that conversants act depending upon the context which reflects some characteristics of the on-going speech but is not necessarily seen as embodying meta-level messages, and that conversational organization emerges from these actions of the speaker and the hearer (see the lower part of Figure 1.1). Such contexts include external events surrounding conversants as well as conversants' behaviors at the time the conversation is taking place.

In Chapter 5, we try to develop a non-signaling model of turn-taking. We discuss how smooth transitions of speaking turn can be achieved without assuming the encode/decode of turn-taking signals.

1.4 Outline of the Dissertation

The rest of the dissertation is organized as follows.

In Chapter 2, we present the dialogue corpus data used throughout the dissertation with a detailed description of procedures for data collection and transcription and with a basic statistics of the corpus data.

In Chapter 3, we start with exploring the global structural elements, discourse structures. We investigate the relationship between discourse structures and dynamic speech rates in several conditions, elucidating the informational potentials of dynamic speech rate in dialogues. We claim that changes in the speech rate have definite potential in signaling the structure of information collaboratively constructed by the conversants of a conversation.

In Chapter 4, we pay attention to the local structural elements, turn-taking and backchannels. We explore the predictive powers of prosodic and syntactic features of the speaker's speech for discriminating turn-taking categories and backchannel categories, elucidating the function of these features as contextualization cues to turn-taking and backchannels. We further discuss the interrelationship between prosody and syntax in the decision process of turn-taking and backchannels.

In Chapter 5, following the descriptive studies presented in the previous two chapters, we develop a model of a conversant's action which tries to understand the mechanism underlying turn-taking phenomena, without supposing direct use of a signaling system. We show that our non-signaling model can account for smooth transitions of turn as well as the regularities observed in wider range of conversational interactions including simultaneous starts of talks and unusual lapses between talks.

Finally, we summarize the dissertation and describe future directions in Chapter 6.

Chapter 2

Japanese Map Task Corpus

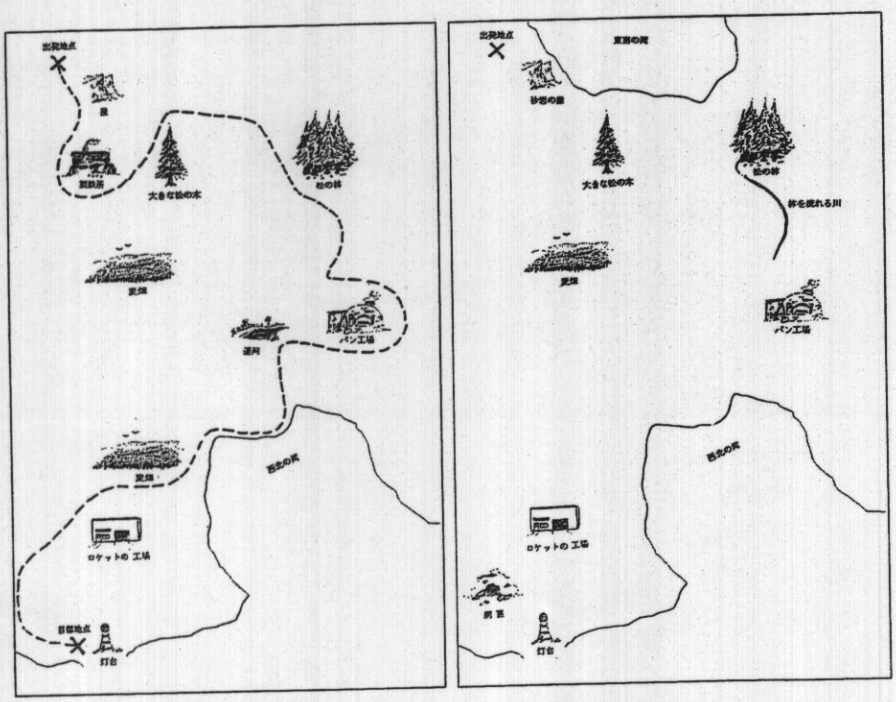
Before moving on to the main task, we explain the dialogue data used throughout the dissertation. In this chapter, we describe procedures for data collection and for transcription, and provide basic statistics of the data.

2.1 Procedures for Data Collection

All of the dialogues used for the analysis were taken from the Japanese Map Task Corpus (Aono et al., 1994; Horiuchi et al., 1997), a Japanese-language version of the Edinburgh Map Task Corpus (Anderson et al., 1991). The basic design follows that of the Edinburgh Map Task Corpus with respect to map and route designs and situational conditions such as familiarity and eye contact.

2.1.1 Task

In the Map Task dialogues, two participants exchange utterances spontaneously and naturally, induced by the design of the task, in which they look at similar but significantly different maps of the same region, each unseen by the other. The party with the route (the "Instruction Giver") instructs the other (the "Instruction Follower") to draw a route through landmarks on the map from the start to the goal (see Figure 2.1 for sample maps).



Giver's map Follower's map

Figure 2.1: Sample maps used in the Map Task dialogues.

2.1.2 Materials

2.1.2.1 Feature Types

All map routes begin with a starting point and end with a finishing point; starting points are marked in the same way on the Instruction Giver's and the Instruction Follower's maps, while finishing points are marked only on the Giver's maps. Intermediate landmarks along the Giver's route alternate between those that are common to the Giver's and Follower's maps, and those that differ. Each map contains each of the following types, which are all not common to both maps.

Absent/Present landmarks which are found on the Giver's map but not on the Follower's, and vice versa.

Name Change landmarks which have identical forms and locations but different names on the two maps.

2:1 landmarks which appear twice on the Giver's map, once in a position close to the route and once far away. The Follower has only the one far away from the route.

2.1.2.2 Routes

A route is drown around the landmarks observing the following criteria:

- The route starts at landmarks common to both maps;
- The route finishes at a common landmarks; and
- Intermediate landmarks along the route alternate between common features and those that differ in some way.

2.1.2.3 Feature Names

The feature names were carefully selected to highlight linguistically and phonologically interesting phenomena in unsolicited oral discourse in Japanese, such as nasalization (e.g., "ginkoo," "kinkoo") and semivowel-deletion (e.g., "zyuzyutu," "mazyutu").

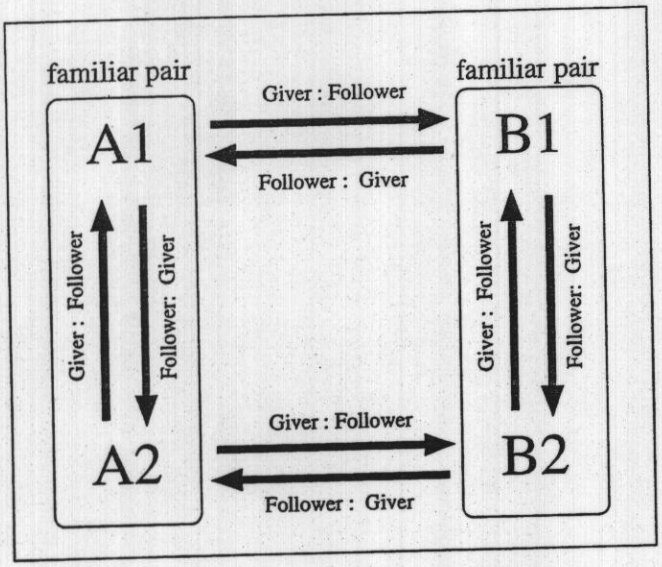


Figure 2.2: Design of experiments. A1 is familiar with A2, and B1 with B2; while A1, A2 are unfamiliar with B1, B2.

2.1.3 Subjects

Sixty-four undergraduates at Chiba University (32 male and 32 female) took part. The subjects' ages ranged from 18 to 24.

2.1.4 Design

Each subject was recruited with a familiar partner of the same sex who know him/her well and was mixed with another pair of subjects of the same sex who were unfamiliar to him/her. Thus, the two pairs formed a quadruple of subjects. Every subject participated in the Map Task four times, twice as Instruction Giver, twice as Instruction Follower, once in each case with his or her familiar partner, once with an unfamiliar partner (Figure 2.2). They used a different set of maps in each session.

Of sixteen quadruples, the half performed all four tasks while able to see the

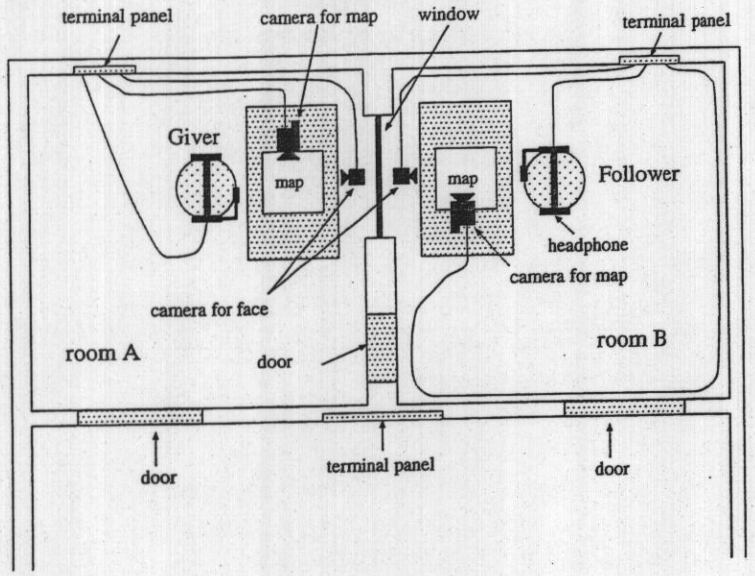


Figure 2.3: The recording environment.

other participant's face ("eye contact condition"), the other half while unable to do so ("no-eye contact condition"); of eight quadruples of each condition, the half were male, the other half were female. The design shown in Figure 2.2 was used in both eye contact and no-eye contact conditions.

2.1.5 Procedures

The speakers were seated in two acoustically insulated rooms (Figure 2.3), where their speech was recorded separately in left-right channels of DAT tapes. In the eye contact condition, the speakers saw each other's upper body through a glass window on the wall, while the window was covered in the no-eye contact condition. The speakers were told (1) that the goal of the task was to draw the Giver's route on the Follower's map, (2) that the Giver's and Follower's maps might be different in some respects, and (3) that neither could use gestures.

2.2 Dialogue Data

The dialogue data consists of the digital recordings of 128 dialogues (totaling approximately 22 hours). The speech data was sampled at 20KHz and with 16-bit accuracy, and transcribed in Japanese.

2.2.1 Unit for Description and Analysis

In this section, we describe the unit for description and analysis. Although the concept of the unit of a dialogue or conversation is essential both for description and analysis, there is no single accepted definition for the unit. For example, grammatical units can be used as a unit of description and analysis. They, however, seem inadequate for spontaneous conversations because in such conversations there are various obstacles to grammatical constructions, such as repairs, hesitations, and interruptions by other speakers, which make the judgment of grammatical units considerably difficult. Speaking turns can also be the unit of description and analysis. The concept of 'turn,' however, has been defined by many researchers in many different ways (Sacks et al., 1974; Edelsky, 1981). Thus, the identification of turns in real conversations is also very difficult (Goodwin, 1981).

More objective definitions of the unit of analysis have been provided by some researchers, who have employed pauses as delimiters (Maynard, 1989; Koopmans-van Beinum & van Donzel, 1996; Nakajima & Tsukada, 1997). Pauses are physically detectable as particular regions in speech based on energy measurements, provided that high-quality recorded speech materials are available. This resolves difficulty in objectively and reliably defining the unit. The unit defined in terms of pauses is also useful for the analysis of turn-taking, since in Japanese conversations, as Maynard (1989) pointed out, turn changes occur most frequently at or during pauses. For these two reasons, we define our unit of description and analysis in terms of pauses. We call such a unit an *inter-pausal unit* (IPU), which is a stretch of a single speaker's speech bounded by pauses longer than 100 msec. The speech signals were automatically divided into IPUs using energy

measurements.[1]

Before leaving the discussion on the unit of description and analysis, let us mention other phonetically defined units such as intermediate and intonational phrases (Pierrehumbert & Hirschberg, 1990), which have often been used by researchers in the field of discourse and dialogue (Hirschberg & Nakatani, 1996). Identification of those phrases is based on detection of extra melodic elements such as phrase accent and boundary tone, phrase-final syllable lengthening, and pausing (Pierrehumbert & Hirschberg, 1990); this requires hand-labeling by experts.[2] Inter-pausal units, on the other hand, can be extracted automatically. Therefore, we prefer to use IPUs.

2.2.2 Transcription

The speech data was transcribed in the Japanese standard orthography based on inter-pausal units. The information contained in the Japanese Map Task corpus is illustrated in the following excerpt taken from the corpus (Table 2.1). Each line corresponds to an IPU. The second and third columns, each having a minute:second:millisecond format, are the start time and end time of the IPU, respectively. The rightmost column contains transcriptions in Japanese of what was said. "G" stands for the Instruction Giver and "F" the Instruction Follower.

2.3 Basic Statistics of the Data

In this section, we provide basic statistics of the data. In particular, we see the influence of the individual roles, i.e., the Giver and the Follower, on the temporal aspects of conversations. Here, we examine eight dialogues, whose portion we will make use of in the following chapters.

[1] Koopmans-van Beinum and van Donzel (1996) list silent pauses, filled pauses, and the lengthening of certain words as 'pauses,' but in this study, we regard only silent pauses as those defining IPUs. We, however, exclude silent pauses caused by plosives, fricatives, and geminate consonants appearing within words or between content words and function words.

[2] Some researchers have recently started to work on the automatic labeling of intermediate and intonational phrases (Wightman & Ostendorf, 1994; Campbell, 1996), which would make it easier for us to label speech data on the basis of such units.

Table 2.1: An excerpt from the Japanese Map Task Corpus.

	Start time	End time	Transcript
1	00:01:888	00:02:464	G: dewa iidesuka (are you all right?)
2	00:02:528	00:02:704	F: hai (yes)
3	00:03:184	00:04:640	G: zya mazu syuppatutitenni imasuyone (you are at the starting point, aren't you?)
4	00:04:704	00:04:864	F: hai (yes)
5	00:05:376	00:05:872	G: etto- (uh)
6	00:06:240	00:06:624	G: zya mazu (first of all)
7	00:06:784	00:07:696	G: sitani mukatte (in the direction going down)
8	00:07:840	00:08:636	G: syuppatusite kudasai (please start)
9	00:08:528	00:10:288	F: e-to- sitato iunowa (uh, in the direction going down)
10	00:10:464	00:11:248	F: e-to masitade (uh, right below?)
11	00:11:488	00:12:112	F: yorosiidesuka (is it right?)
12	00:12:320	00:13:792	G: a tyotto hidarini itte sitani (go a little left and down)

Table 2.2: Whole duration of each dialogue and total duration of IPUs for each role (minute:second:millisecond).

Dialogue	Whole duration	Duration of IPUs	
		Giver	Follower
1	07:03:120	03:35:828	02:00:120
2	08:04:800	03:38:480	01:11:336
3	08:31:808	03:39:549	02:09:180
4	11:54:240	04:32:962	02:55:563
5	06:41:904	02:31:824	01:47:056
6	10:12:624	04:58:944	02:02:176
7	09:00:656	04:16:748	01:29:420
8	10:42:576	04:52:156	02:48:128
Average	09:01:466	04:00.811	02:02:872

Table 2.2 shows the whole durations of each dialogue as well as the total duration of IPUs excluding pauses, for each role. Table 2.3 shows the number and the mean duration of IPUs. It is observed that the speech duration of the Giver is generally longer than that of the Follower. This difference is due to the differences both in the number of IPUs and in the mean duration of IPUs; the Giver tends to produce more IPUs, each of which, in general, has longer duration compared to these of the Follower (see Table 2.3). Thus, we have to give careful consideration to such differences between the roles when the temporal aspects of speech are discussed in Chapter 3.

Table 2.3: Mean duration of IPUs for each dialogue (msec).

	All			Giver			Follower		
Dialogue	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
1	437	768.8	648.0	252	856.5	602.3	185	649.3	689.3
2	369	785.4	628.8	219	997.6	674.2	150	475.6	385.7
3	356	979.6	835.5	180	1219.7	876.7	176	734.0	714.1
4	597	753.0	581.8	369	737.3	555.4	226	777.0	623.1
5	320	809.0	652.1	164	925.8	650.1	156	686.3	633.4
6	561	750.7	664.8	299	999.8	716.6	262	466.3	457.9
7	491	705.0	608.4	317	809.9	637.5	174	513.9	499.2
8	526	875.1	696.8	306	954.8	664.7	220	764.2	726.3
Total	3657	795.9	665.6	2107	914.4	676.4	1550	634.7	615.3

Chapter 3

The Relation of Dynamic Speech Rate to Discourse Structure

In this chapter, we start by exploring the global structural elements, *discourse structures*. We investigate the relationship between discourse structure and dynamic speech rate in several conditions, elucidating the informational potentials of dynamic speech rate in dialogues. We claim that changes in the speech rate have potential in cuing the structure of information collaboratively constructed by the conversants.

3.1 Introduction

There are a number of phenomena which are strongly influenced by the behavior of the speaker and of the hearer in conversational interaction. The investigation of such phenomena helps not only to make clear the contexts of these phenomena, but also to elucidate potentials of exchanging meta-level messages based on the contexts. In this chapter, we focus on the global structural elements, discourse structures, exploring such exchanges of meta-level messages with respect to discourse structure.

It has been supposed that a discourse can be partitioned into segments, and each segment can embed and be embedded in other segment, resulting in a hierarchical structure (Grosz et al., 1989). Thus, the structure of a discourse is very

complicated, and therefore the conversants can be supposed to use various clues for understanding the structure of a discourse.

Various prosodic features, such as pitch range, intonation, pause, and duration, have been investigated from the perspective of their potentials as contextualization cues to the structure of a discourse, while not much research has been devoted to speech rate phenomena. This is partly because the primary interest in speech rate phenomena has been directed toward the apparent rhythmicity in speech and its relationship with underlying physiological mechanisms of speech production and perception.

Speech researchers have been studying the temporal organization of human speech and some have suggested that an isochronous unit can be hypothesized behind speech production and perception (Lehiste, 1977; Marcus, 1981), even though the actual speech materials do not manifest exact isochrony due to intervening physiological and linguistic factors. It has also been observed among speech engineers (Sagisaka & Tôkura, 1984; Fujisaki & Higuchi, 1980) that control of the temporal organization of synthetic speech greatly contributes to its naturalness. Their interests have also been focused on physiological and cognitive mechanisms of human speech production and perception, though their results are mostly based on read speech data of at most sentence length.

In contrast, the present study focuses on the informational potential of speech rate changes observed in human spontaneous dialogues. In conversations, we notice that people change their rates of speech from time to time, and that these local modulations in the speech rate, together with insertions of pauses, provide the hearer with various clues about the state of the speaker and the state of the ongoing conversations, including the structure of a discourse. Therefore, deviations from an isochronous speech rate in conversations have the potential to be contextualization cues to discourse structure. In this study, we focus our attention on local changes of the speech rate, namely, *accelerations* and *decelerations* of speech over consecutive pairs of utterances.

Several researchers have reported that speech tends to be slow in the initial parts of paragraphs, sentences, and discourse segments and tends to be fast in the final parts (Brubaker, 1972; Koopmans-van Beinum & van Donzel, 1996; Hirschberg & Nakatani, 1996). These studies are, however, all concerned with

some forms of *monologues*, and very few attempts have been made at *dialogues*. Dialogue has multiple speakers, bringing about various interactive aspects such as turn-taking and backchannels, and therefore, some questions that have never been posed in monologue research arise; whether acceleration patterns are universal across single-speaker utterances and cross-speaker utterances; whether the insertion of backchannels by other parties change the speech rate characteristics within a single speaker's utterances. Thus, the relation between changes of speech rate and discourse structure in dialogues has left various unresolved problems.

Swerts and Ostendorf (1997), who, to the best of my knowledge, conducted the only study on the relationship between speech rate and information structure in *dialogues*, found no correlation between discourse structure and speech rate which has been observed in monologues. Their target was, however, not dialogues among humans but those between humans and computers. Possible reasons for their results to differ from previous ones may be, as they pointed out, the unnaturalness of human-machine conversations and the differences between the monologue settings in other studies and the dialogue settings in their own studies. Accordingly, we should investigate the behaviors of speech rates in dialogues conducted by humans only, examining whether, in dialogues, there is the same relationship between discourse structure and speech rate as in monologues.

In the first half of this chapter, we concentrate on examining the regularities we have found in our dialogue corpus between speech rate and information structure in spontaneous dialogues. After introducing the method in which the corpus data is analyzed, we report on the basic results of our analyses. We examine relevant features, both individual and task specific, to establish that the regularities we have found are significant. The second half of the chapter is devoted to recasting our findings in terms of informational signaling or cuing by dynamic speech rate. We propose an analysis of the signaling potential making use of performance measures developed in information retrieval research. Finally, we elaborate and enforce our findings by showing that the observed cuing by dynamic speech rate are not an spurious one induced by competing cues to discourse structure.

3.2 Methods

3.2.1 Materials

We examined four task-oriented dialogues totaling 40 minutes conducted by different pairs of speakers. All of the dialogues were taken from the Japanese Map Task Corpus (see Chapter 2). The dialogues we analyzed included all possible combinations of familiarity and facing conditions.

3.2.2 Speech Rate Measurement

It has been argued that the temporal organization of speech can be classified broadly into two categories (Pike, 1945; Abercrombie, 1967). One is the "stress-timed rhythm" in which stressed syllables tend to be produced at isochronous intervals. Languages such as English, Dutch, and German belong to this type. The other is the "syllable-timed rhythm" which is organized by keeping the durations of all syllables approximately equal. Languages such as French, Italian, and Spanish belong to this type. Japanese, in which rhythmic patterns are organized not by syllables but by morae, belongs to the syllable-timed rhythm or "mora-timed rhythm" category. Morae in Japanese tend to keep their durations equal by expanding or contracting phonemes (composing the morae) according to their adjacent phonemes (Sagisaka & Tôkura, 1984). Therefore, we consider the duration of a mora as one of the standards against which we can measure speech rates for Japanese.[1]

Our measurement of the speech rate is based on the average mora duration of *inter-pausal units* (IPU) (see Section 2.2.1). The *average mora duration* (AMD) of an IPU is the duration of the IPU divided by the number of morae appearing in the IPU. Therefore, a smaller AMD of an IPU means a higher speech rate over the IPU, and a larger AMD means a lower speech rate.

In order to capture the *changes* of the speech rate in dialogues, we compare

[1]On the other hand, there is a large variability in an average syllable duration in stress-timed languages such as English and Dutch, and this aspect of syllables complicates speech rate measurements based on syllables (Koopmans-van Beinum & van Donzel, 1996).

Table 3.1: An excerpt from the Japanese Map Task Corpus with AMD for each IPU (msec).

	Start time	End time	Duration	# morae	AMD	Transcript
1	01:47:552	01:49:152	1600	10	160.0	G: tugini ookina iwato
2	01:49:344	01:49:488	144	2	72.0	F: hai
3	01:49:488	01:51:312	1824	21	86.9	G: indianno murano aidao tootte kudasai
4	01:51:984	01:52:144	160	2	80.0	F: hai
5	01:54:432	01:54:848	416	3	138.7	G: tugini
6	01:55:024	01:56:112	1088	8	136.0	G: e ookina iwato
7	01:57:280	01:58:352	1072	10	107.2	G: kinno koozandesuka
8	01:58:480	02:00:080	1600	20	80.0	G: kinno koozanno aidao tootte kudasai

the speech rates of a pair of temporally consecutive IPUs.[2] We call the pair *accelerating* if the AMD of the succeeding IPU is smaller than that of the preceding IPU; the pair is *decelerating* if the AMD of the succeeding IPU is greater than that of the preceding IPU. See, for example, the excerpt from our transcription in Table 3.1. The AMDs of IPU1 and IPU2 are, respectively, 160.0 msec and 72.0 msec. This makes the AMD of IPU2 smaller than that of IPU1 by 88.0 msec, and hence the IPU pair $\langle 1,2 \rangle$ is classified as an accelerating pair. On the other hand, the AMD of IPU5 is greater than that of IPU4 by 58.7 msec, and hence the IPU pair $\langle 4,5 \rangle$ makes a decelerating pair. Likewise, the IPU pairs $\langle 3,4 \rangle$, $\langle 5,6 \rangle$, $\langle 6,7 \rangle$, and $\langle 7,8 \rangle$ are accelerating.

We choose the bare AMD of an IPU as the measure of the speech rate. One might argue that we could have taken the variability in the lengths of phonemes comprising the morae into account (Sagisaka & Tôkura, 1984). In addition, we could have used a normalized mora duration, instead of a bare AMD, based on the mean and the standard deviation of the length of each phoneme occuring in the mora (Campbell & Isard, 1991). We, however, mainly make use of a bare AMD as the measure of the speech rate, because of the following reasons. Firstly, the

[2]We exclude those IPUs that overlap other IPUs made by different speakers temporally, which were uttered in a low voice like murmuring, and which contain hesitations within words.

status of the normalization of phoneme lengths not taking the adjacent phonemes into account is questionable, since in Japanese, the durations of phonemes tend to expand or contract according to their neighboring phonemes (Sagisaka & Tôkura, 1984). Secondly, it has not yet been established that our perception of the speech rate is based on normalized mora durations rather than mora durations. If variations in phoneme durations do contribute to our perception of the speech rate, it is better not to obfuscate that effect through normalization. Thirdly, we do not have enough data to calculate normalized phoneme durations of our speakers reliably. The comprehensive phoneme length data currently available are taken from read speech, which is not suitable for application to our dialogue data.

However, a bare AMD is not surely the best measure of the speech rate. Thus, as a supplementary analysis to confirm the results based on bare AMDs, we also make use of an average mora duration normalized by the mean length of each phoneme which is taken from read speech.

There are specific circumstances in which longer than normal mora durations are regularly produced. These circumstances include the lengthening of final syllables, the occurrence of filled pauses, and short IPUs, typically consisting of one or two morae. Rather than discarding these instances as anomalies, we have included them in our analyses, even though these instances can cause a larger variability in the entire AMD range. We do not want to exclude these instances from the beginning, because they may also serve to contribute to the phenomena we are looking at. We will discuss some of these exceptional phenomena in later sections.

3.2.3 Informational Structures

We intend to study whether the changes of speech rate can function as contextualization cues to the openings of *information units* in dialogues. For this purpose, we classify pairs of IPUs into *opening pairs* whose second IPUs start the presentation (typically new) pieces of information and *non-opening pairs* whose second IPUs do not start such presentation. For example, among the pairs of consecutive IPUs in Table 3.2, the pair $\langle 5, 6 \rangle$ is classified as a non-opening pair on this scheme, while the pair $\langle 4, 5 \rangle$ is classified as an opening pair. Note that the non-

Table 3.2: Acceleration/deceleration and opening/non-opening properties of pairs of IPUs.

	1	160.0	G: tugini ookina iwato	
accel			(then, between a big rock and)	non-open
	2	72.0	F: hai	
accel			(yeah)	non-open
	3	86.9	G: indianno murano aidao tootte kudasai	
accel			(the Indians' village please go through)	non-open
	4	80.0	F: hai	
decel			(right)	open
	5	138.7	G: tugini	
accel			(and then)	non-open
	6	136.0	G: e ookina iwato	
accel			(ah, between the big rock and)	non-open
	7	107.2	G: kinno koozandesuka	
accel			(a gold mine, isn't it?)	non-open
	8	80.0	G: kinno koozanno aidao tootte kudasai	
			(and a gold mine please go through)	

opening pair $\langle 5,6 \rangle$ is an accelerating pair according to the speech rate measure, and the opening pair $\langle 4,5 \rangle$ is a decelerating pair. Our task is to study these kinds of possible correlations between the acceleration/deceleration properties of the pairs of IPUs and their opening/non-opening properties.

Note that our criterion for opening/non-opening is simply whether the second IPU of a pair starts a presentation of information, and it is conceptually different from whether the first IPU *ends* a presentation of information or whether the second IPU *continues* the presentation of information that the first IPU contributes to. Therefore, our notion of opening is close to the initiation of full or subordinate "discourse units" in David Traum's theory of grounding (Traum, 1994). Our demarcations, however, deviate from his, where the response to a request or

the answer to a query is considered as completing the expression of information initiated by the request or question part. This is to respect the intuition that the adjacency pair of request-response or query-answer expresses a single piece of information to be shared by the conversants, concerning an issue raised or an action requested. In Traum's theory, however, a response or an answer is considered to initiate an independent discourse unit.

The actual labeling of opening/non-opening was a "consensus labeling" by two labelers, who examined the transcribed text and the recorded speech of dialogues until they reached an agreement on their judgments. The task for the Map Task dialogues consists of subtasks for the identification of landmarks and the tracing of paths by connecting the landmarks. Thus, the structure of the task itself makes clear what information is to be exchanged and in what order. Therefore, it is relatively easy to reach an agreement on when a piece of information is introduced during the dialogue. This is important because we need an independent characterization of discourse structures in studying the relationship between dynamic speech rates and discourse structures to avoid circular arguments.

At the same time, since our data consists of spontaneous dialogues, we have to make some non-trivial decisions on how to deal with the dialogue-specific forms of information exchanges typically found in such data. The following gives the guidelines that we have adopted:

Backchannels Whenever an IPU pair has a backchannel as its second IPU, we label the pair as non-opening. A backchannel is a short utterance, such as "yeah" or "uh huh" in English uttered by the hearer without claiming the shift of turn (Maynard, 1989; Schegloff, 1982).[3] Whatever the exact functions of backchannels may be, it is clear that they contribute no information to the topics of dialogues, and hence do not start a presentation of information.

Backchannel flanking Since a backchannel has no information content about the topics of dialogues, it makes little sense to consider informational con-

[3]We judged backchannels from their forms and functions. Expressions such as "hai," "ee," and "un" in Japanese were judged to be backchannels unless they constituted grounding acts such as an answer to a yes-no question (Traum, 1994).

tinuity of a backchannel and the succeeding IPU. Accordingly, to judge whether such a succeeding IPU opens a presentation of information or continues the preceding one, we compare the IPU with the IPU *preceding* the backchannel, rather than the backchannel itself. That is, instead of labeling IPU pairs whose first IPUs are backchannels, we label IPU pairs *flanking backchannels.*

Conjunction When an opening of information is headed by a conjunctive expression such as "sorekara (and then)" or "sorede (therefore)," we consider the information to open at the beginning of the conjunctive expression.

Restarts When a speaker restarts a presentation of information after a false start or the cancelation of a previous attempt at presenting the information, we regard both the false start and the restart as openings of information.

This labeling results in a 4.0 sec average distance (SD 2.8 sec) and 3.7 IPUs (SD 2.1) between the opening of information and the next opening of information. We call a sequence of IPUs bounded in this way an *information unit.* There are a few cases where an opening of information occurs in the middle of an IPU rather than at an IPU boundary, but they account for only 5% of all instances of information openings, so we decided to ignore them.

3.2.4 Types of Inter-Pausal Unit Pairs

We classify pairs of IPUs into four categories in terms of conversational turn structures. The categories are:

(A) the pairs of directly adjacent IPUs made by the same speaker. E.g., the IPU pairs $\langle 5,6\rangle$, $\langle 6,7\rangle$, $\langle 7,8\rangle$ in Table 3.2.

(B) the pairs of IPUs made by the same speaker flanking a backchannel. E.g., $\langle 1,3\rangle$.

(C) the pairs of IPUs involving turn changes between speakers. E.g., $\langle 3,4\rangle$, $\langle 4,5\rangle$.

(D) the pairs of IPUs involving voiced backchannels from the hearer but no turn changes. E.g., $\langle 1,2\rangle$.

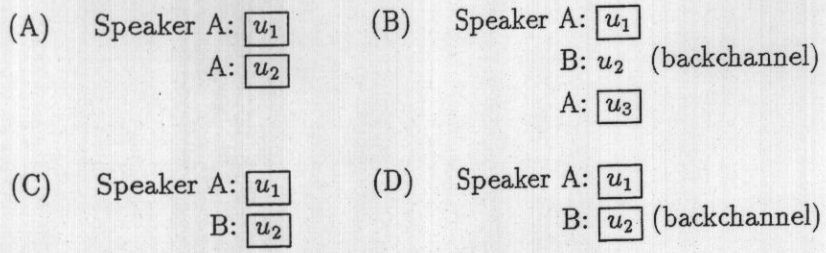


Figure 3.1: Categorizations of IPU pairs based on turn structures.

The temporal structures of the pairs for each of these categories are shown in Figure 3.1. For the reason explained above, we have category (B), instead of a category of IPU pairs whose first IPUs are backchannels. Also, note that IPU pairs of type (D) are invariably labeled as "non-opening."

3.3 Results

3.3.1 Preliminary Study

Table 3.3 shows the mean AMDs of the individual conversants of our target dialogues, as well as the frequencies of information openings made by them. ("1G" denotes the speaker who played the role of Giver in dialogue 1, while "2F" denotes the speaker who played the role of Follower in dialogue 2, and so forth.)

Here, we can find some differences in the average AMD over different speakers. The mean AMD of a speaker who plays the role of Giver tends to be greater than that of a speaker who plays the role of Follower. In fact, when we examine dialogues in which the same pairs of speakers play reverse roles, we find that the relationship between those speakers' mean AMDs is also the reverse (see the role-reverse data in Table 3.3). Therefore, it can be concluded that the average AMD seems to be more strongly influenced by the difference in roles than by the difference in individual speakers.

Table 3.3 also shows that Givers tend to open information units more frequently than Followers, suggesting that the average AMD of a speaker might

Table 3.3: Frequencies of openings of information units and mean AMDs of individual speakers (msec). AMDs of the same speakers in role-reversed dialogues are also shown for comparison.

	Target data				Role-reverse data		
	# of		AMD			AMD	
Speaker	IU openings	Role	Mean	SD	Role	Mean	SD
1G	67	Giver	127.8	39.9	Follower	103.3	28.3
1F	29	Follower	100.8	32.6	Giver	134.5	56.2
2G	88	Giver	140.5	73.6	Follower	96.6	29.0
2F	32	Follower	112.0	42.5	Giver	144.9	76.5
3G	67	Giver	139.0	72.2	Follower	112.5	29.0
3F	31	Follower	106.1	44.8	Giver	124.4	39.0
4G	48	Giver	130.5	44.7	Follower	122.2	40.3
4F	33	Follower	113.3	44.1	Giver	130.3	47.8

ultimately be affected by the frequencies of information openings made by him or her. This motivates us to take a closer look at the relationship between information openings and speech rates.

3.3.2 Speech Rates over Information Units

We compared average speech rates of the initial, intermediate, and final parts of information units. More precisely, we classified all IPUs in our data into *starting* IPUs at the beginning of information units, *ending* IPUs at the end of information units, and *intermediate* IPUs (i.e., neither starting IPUs nor ending IPUs).[4] We then calculated the mean AMDs for each class of IPUs.

Table 3.4 shows that the rate of speech significantly differs depending on where in an information unit the speech is located ($F(2,14) = 43.6$, $p < .01$). Multiple comparison tests by Newman-Keuls procedure show significant differences in all

[4]When an information unit consists of a single IPU, that IPU is both a starting IPU and an ending IPU.

Table 3.4: Mean AMDs of starting, intermediate, and ending IPUs (msec).

	All			Giver			Follower		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Starting	395	143.6	60.3	270	148.7	65.2	125	132.6	43.1
Intermediate	949	114.0	38.6	513	124.8	43.6	436	101.3	26.8
Ending	395	103.9	33.9	216	109.2	31.1	179	97.6	36.1

pairwise comparisons ($p < .05$ comparing the intermediate and ending IPUs, and $p < .01$ for any other comparison).

Table 3.4 also shows that the speech rates significantly differ in each case of the Giver and of the Follower (Giver: $F(2,6) = 14.8, p < .05$; Follower: $F(2,6) = 86.5, p < .01$). Multiple comparisons show significant differences in all pairwise comparisons with the exception of the comparison between the intermediate and ending IPUs in the case of the Follower ($p < .05$ comparing the intermediate and ending IPUs for the Giver, and $p < .01$ for any other comparison). These results mean that speech is relatively slow at the beginning of an information unit and is fast at the the end of the information unit, and that the speech rate shows this tendency regardless of the role of the speaker, the Giver and the Follower.

As we mentioned before, Hirschberg and Nakatani (1996) made a number of observations about the relationship between speech rates and information structures in read and spontaneous monologues, including the claim that speech rates tend to be high toward the end of a discourse segment. Our results are parallel to their findings, although ours are concerned with *information units* in *dialogues* as opposed to discourse segments in monologues, which are generally larger than our units.

3.3.3 Local Changes of Speech Rates

Figures 3.2 (A) to (D) respectively show the distributions of IPU pairs of categories (A) to (D), relative to their properties of the acceleration/deceleration and opening/non-opening of information units. The abscissa of each figure marks

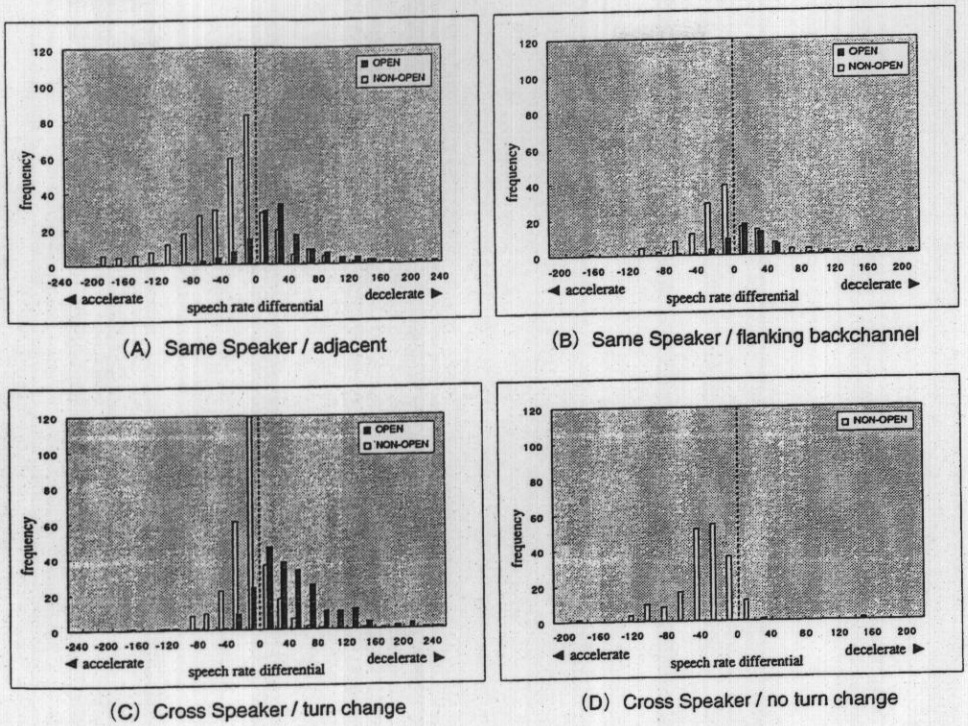


Figure 3.2: Distribution of the local speech rate changes for each turn structure condition. (The black and white columns over a range represent the number of opening pairs and the number of non-opening pairs, respectively, falling within the range.)

the range of local speech rate changes over the IPU pair. Since backchannels, by definition, do not start new information units, all of the data in Figure 3.2 (D) are classified as non-openings. These figures clearly show that there are distributional differences between opening and non-opening IPU pairs in local dynamic speech rate changes. The area to the left of the zero line has much more (and higher) white columns than black columns, showing that the majority of the accelerating pairs of IPUs are non-opening pairs. On the other hand, the area to the right of the zero line has much more black columns, indicating that the majority of the decelerating pairs of IPUs are opening pairs. Therefore, we can say that openings of new information units tend to co-occur with decelerations of speech,

Table 3.5: Variations in average speech rates by individual dialogues. The mean AMDs are shown for each dialogue (msec).

Condition		Dialogue 1	2	3	4	Total
(A)	open	23.9	24.0	22.6	31.1	25.2
	non-open	−22.8	−57.9	−40.3	−26.4	−35.3
(B)	open	48.2	67.7	4.9	18.3	40.9
	non-open	8.9	−12.0	−1.3	−22.3	−2.5
(C)	open	30.9	62.4	45.3	55.1	48.6
	non-open	−15.3	−4.9	−11.6	−19.6	−11.8
(D)	non-open	−44.3	−40.3	45.9	−28.3	−40.1

and non-openings tend to co-occur with accelerations of speech. Note that the co-occurrences are also universal across IPU pairs in different turn structures. Table 3.5 shows that these tendencies are fairly stable across individual dialogues.

We applied a χ^2 test to confirm the observations made so far. The test was to measure the strengths of the associations between the acceleration/deceleration properties of IPU pairs and their opening/non-opening properties. Table 3.6 shows the frequencies of accelerating/decelerating and opening/non-opening IPU pairs for each of the four conditions (A) to (D). As expected, we found relatively strong associations between these parameters for the entire class of IPU pairs ($\chi^2(1) = 513.3$, $p < .01$, $\phi = 0.62$) and within the individual categories of IPU pairs ((A): $\chi^2(1) = 143.9$, $p < .01$, $\phi = 0.56$; (B): $\chi^2(1) = 35.4$, $p < .01$, $\phi = 0.44$; and (C): $\chi^2(1) = 215.3$, $p < .01$, $\phi = 0.66$).[5]

There are two points that need to be noted. First, from Figures 3.2 (A) and (B), it is evident that the insertion of backchannels by other parties does not change the speech rate characteristics within a single speaker's utterances.

[5]The *coefficient of association* ϕ for data in categorical values is the counter-part of the coefficient of correlation r for data in continuous values. ϕ takes real values between 0 and 1; 1 means the maximal association between two categories, whereas 0 means no association. Note that a χ^2 test is not applicable to IPU pairs of type (D), since there are no opening pairs of IPUs of that type.

Table 3.6: Frequencies of accelerating/decelerating and opening/non-opening IPU pairs.

	non-open		open	
Condition	accel	decel	accel	decel
(A)	265	64	27	105
(B)	89	42	9	42
(C)	227	62	25	187
(D)	192	8	–	–
Total	773	176	61	334

Second, a comparison of Figures 3.2 (A)(B) with Figures 3.2 (C)(D) shows that local speech rate characteristics are universal across single-speaker utterances and cross-speaker utterances.

3.3.4 Role Differences

As we mentioned before, the average AMD of the Giver tends to be greater than that of the Follower. Given our results about dynamic speech rates and information openings, this might be only natural. We found that the Giver tends to have IPUs at the beginning of information units, and that IPUs are slower at the beginning of an information unit than at the the end of the information unit. Consequently, the average AMD of the Giver might be greater than that of the Follower.

One might, however, argue the opposite. The Giver, by some reason, might speak more slowly than the Follower. Moreover, since the Giver often takes the initiative to give instructions, the Giver opens information more often, while the Follower closes information more often. Then, *on the surface*, speech is relatively slow at the beginning of an information unit and is fast at the the end of the information unit, although there is no direct correlation between information openings and speech deceleration. In this theory, the apparent correlation is due to the intrinsic fastness of the Follower's speech, the intrinsic slowness of the Giver's speech, and the fact that most information openings are made by the

Table 3.7: Frequencies of accelerating/decelerating and opening/non-opening IPU pairs for conditions (C) and (D) elaborated by the role differences. "G→F" means the case in which the IPUs of the Giver precede the Follower's, and "F→G" means the opposite case.

Condition		non-open		open	
		accel	decel	accel	decel
(C)	G→F	142	26	14	65
	F→G	85	36	11	122
(D)	G→F	170	5	–	–
	F→G	22	3	–	–

Giver after the Follower's ending speech.

To examine the validity of the above argument, we studied whether the local speech rate characteristics that we observed hold in cases where the IPU made by the Giver precedes the IPU made by the Follower and also in the opposite cases (Table 3.7). In both cases, we found strong associations between the acceleration/deceleration properties of IPU pairs and their opening/non-opening properties ((C) G → F: $\chi^2(1) = 100.2$, $p < .01$, $\phi = 0.64$; (C) F → G: $\chi^2(1) = 100.9$, $p < .01$, $\phi = 0.63$). Therefore, the correlation between dynamic speech rates and information openings/non-openings holds *regardless* of the order of the Giver's IPU and the Follower's IPU. It does not depend on the particular order of the Giver's speech and the Follower's speech.[6]

We conclude that there are correspondences between speech accelerations and the absence of information openings, and between speech decelerations and the presence of information openings. This means that speech tends to accelerate until the end of the information unit, and that it tends to decelerate at the be-

[6]A somewhat related question would be whether any effects of role differences are observed in the same speaker conditions. Examination of our data confirms that the correlation between dynamic speech rates and information openings/non-openings holds both in Giver-Giver and in Follower-Follower transitions ((A) G → G: $\chi^2(1) = 91.6$, $p < .01$; (A) F → F: $\chi^2(1) = 49.2$, $p < .01$; (B) G → G: $\chi^2(1) = 25.6$, $p < .01$; (B) F → F: $\chi^2(1) = 9.5$, $p < .01$).

ginning of information units.[7] In addition, the correlations in question hold not only for a single speaker's utterances but also for the sequential utterances of multiple speakers with or without turn changes. These results imply that, intentionally or unintentionally, speakers collaborate with one another to maintain these regularities governing dynamic speech rates and discourse structures. Although there have been some interesting results concerning the relationship between the speech rate and the structure of the information being presented (Brubaker, 1972; Koopmans-van Beinum & van Donzel, 1996; Hirschberg & Nakatani, 1996), most studies have taken monologues as their data, and have concerned themselves with absolute speech rates and discourse segments larger than information units. Our findings are original in showing that there is a correlation between speech rates and discourse structures also in the case of *dialogues*, and that this correlation is between *local changes* of the speech rate and *information units*.

3.4 Signaling Potentials of Dynamic Speech Rate

We have found a significant correlation between dynamic speech rates and openings and non-openings of information in dialogues. This finding suggests that the former may *carry information about* the latter, serving as a *signal* as to when a new piece of information starts and when an old piece continues. We will devote the second half of the chapter to a detailed examination of this possibility.

More specifically, we will consider (1) whether we can ascribe such informational potentials to dynamic speech rates in the first place, (2) what kinds of

[7]The question of why speech generally decelerate at the beginning of information units is outside of the scope of the present study. We, however, conjecture that there is a cognitive reason: it may be that there are heavier cognitive load in planning and producing what to say at the beginning of an information unit compared to the end of an information unit, since you have more choices about what to say at the beginning of an information unit. This might slow down the speech at the beginning. Some people point out that speech decelerates at the beginning of other types of units, such as sentences, paragraphs, discourse segments (Brubaker, 1972; Koopmans-van Beinum & van Donzel, 1996; Hirschberg & Nakatani, 1996), and this may be due to the same reason.

Table 3.8: Information retrieval metrics.

	observed	
classified	open	non-open
open (deceleration)	a	b
non-open (acceleration)	c	d

extra factors may interfere with the signaling in question, and (3) whether it is a genuine signaling relation, rather than a spurious signaling induced by some other real signaling relation. These issues will be addressed in the following three sections in turn.

3.4.1 Evaluation of the Signaling Potentials

To characterize and evaluate the signaling potentials of dynamic speech rates for information structures, we look at our findings from the perspective of information retrieval, following Passonneau and Litman (1997), who applied several information retrieval measures in evaluating discourse segmentation algorithms. More specifically, we compare the signaling of information to the retrieving of information, and characterize signaling potentials in terms of three typical performance measures of information retrieval schemes, namely, *precision*, *recall*, and *error*. These measures give us three different criteria to evaluate the signaling potentials of dynamic speech rates.

Given the information retrieval metrics shown in Table 3.8, these rates are given by the following formulas:

$$\text{Precision rate of openings by deceleration signal} = \frac{a}{a+b}$$

$$\text{Recall rate of openings by deceleration signal} = \frac{a}{a+c}$$

$$\text{Precision rate of non-openings by acceleration signal} = \frac{d}{c+d}$$

$$\text{Recall rate of non-openings by acceleration signal} = \frac{d}{b+d}$$

$$\text{Total error rate by acceleration/deceleration signal} = \frac{b+c}{a+b+c+d}$$

Table 3.9: Precision, recall, and error rates of acceleration and deceleration signaling.

	Precision of acceleration	Recall of acceleration	Precision of deceleration	Recall of deceleration	Total error
(A)	90.8%	80.5%	62.1%	79.5%	19.7%
(B)	90.8%	67.9%	50.0%	82.4%	28.0%
(C)	90.1%	78.5%	75.1%	88.2%	17.4%
(D)	–	96.0%	–	–	–
Total	92.7%	81.5%	65.5%	84.6%	17.6%

The precision rate of deceleration or acceleration signal is a measure of how *accurate* deceleration/acceleration is in cuing openings/non-openings of information, and the recall rate of deceleration/acceleration signal is a measure of how *comprehensive* deceleration/acceleration is in cuing openings/non-openings. Thus, the precision and the recall rates give us criteria to evaluate the signaling potential of each of deceleration and acceleration signals. On the other hand, the total error rate, determined by how often *both* acceleration and deceleration signals incorrectly predict openings and non-openings of information, gives us a criterion for predictive power of dynamic speech rate cuing *as a whole.*

We used our entire data in computing precision, recall, and error rates and did not make the division of data into a training set and a test set, since the aim here is not to evaluate or adjust any model constructed from our data, but to evaluate, based on our *dialogue* data, the relationship between speech rates and discourse structures which has been suggested in the previous studies based on *monologues.*

Table 3.9 shows precision, recall, and error rates in each condition.

3.4.1.1 Acceleration Signaling

In each of Figures 3.2 (A) to (D), we see more (and higher) white columns than black columns in the area to the left of the 0 msec line. More specifically, Table 3.9

shows that 90.8%, 90.8%, and 90.1% of the accelerating pairs of types (A), (B), and (C), respectively, are non-opening pairs. (These numbers are the percentage of white columns out of the total number of columns that have negative AMD values in Figures 3.2 (A)–(C).) Overall, 92.7% of the accelerating pairs are non-opening pairs. These numbers indicate that accelerated speech is a fairly *accurate* signal for the non-openings of information.

Note also that in each of the figures, most of the white columns are to the left of the zero line. In fact, Table 3.9 shows that 80.5% of the non-opening pairs of type (A), 67.9% of the non-opening pairs of type (B), and 78.5% of the non-opening pairs of type (C) are accelerating pairs. (These numbers each give the percentage of negative white columns out of the total number of white columns in Figures 3.2 (A)–(C).) A pair of type (D) has a voiced backchannel as the second IPU, and by definition a backchannel opens no piece of information in dialogues. Therefore, if it is a general tendency that non-opening pairs are accelerating pairs, then a great majority of the pairs of IPUs of type (D) must be accelerating. In fact, our data shows that as much as 96.0% of the pairs of type (D) are accelerating. Overall, 81.5% of the non-opening pairs are accelerating. These numbers indicate that accelerated speech is also a fairly *comprehensive* signal for non-openings of information.

3.4.1.2 Deceleration Signaling

From Figure 3.2 we see more (and higher) black columns than white columns in the area to the right of the 0 msec line, except for Figure 3.2 (B). Indeed, 62.1%, 50.0%, and 75.1% of the decelerating pairs of types (A), (B), and (C) are opening pairs, respectively, with the average of 65.5% of all decelerating pairs being opening pairs. In view of this fact, we can say that the deceleration cues to openings of information are moderately *accurate*.

In each of Figures 3.2 (A) to (C), most of the black columns are to the right of the 0 msec line. Table 3.9 shows that 79.5% of the opening pairs of type (A) decelerate, and the same holds for 82.4% and 88.2% of the opening pairs of types (B) and (C), respectively. These numbers indicate that decelerated speech is a fairly *comprehensive* signal for openings of information.

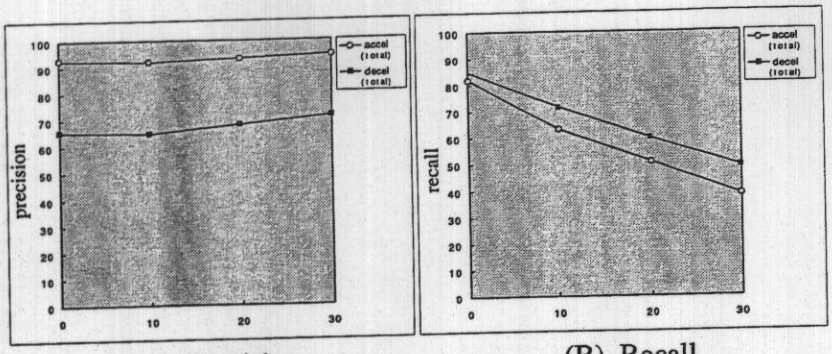
3.4.1.3 Predictive Power of Dynamic Speech Rate Signaling

The total error rates in Table 3.9 show that 19.7%, 28.0%, and 17.4% of the pairs of types (A), (B), and (C), respectively, are incorrectly predicted; in these cases openings and non-openings of information are mispredicted by acceleration and deceleration signals. (These numbers each show the percentage of positive white columns and negative black columns out of the total number of white and black columns in Figures 3.2 (A)–(C).) Overall, 17.6% of the pairs are mispredicted. These numbers are all lower than the chance levels of error rates in the cases of (A), (B), and (C) (40.9%, 40.3%, and 48.8%, respectively). This suggests that a dynamic speech rate can potentially function as contextualization cues for openings and non-openings of information in dialogues.

3.4.1.4 Gray Zone

In the above analysis, acceleration simply means a more than *zero* decrease of AMD over consecutive IPUs, while deceleration means a more than *zero* increase. However, one may well wonder if the boundary between changes of speech rate that signal openings of information and those that signal non-openings can be clear-cut in an actual dialogue. First of all, humans seem to have a certain limitation in their ability to perceive very small changes of speech rates; in order for a change in the speech rate to have any communicative function in a dialogue, it has to be perceived by the dialogue conversants. This gap between the real degrees of speech rate changes and our perception of them is further widened by the fact that our measurement includes several sources of variability such as phrase final lengthening and "native" phoneme durations. Therefore, one may naturally argue that very small degrees of deceleration and acceleration should not have any significant communicative function, constituting a "gray zone" of the signaling to information openings and non-openings.

Unfortunately, perception of changes of the speech rate has not been studied extensively, and hence we have no grounds for assuming the existence of such a gray zone, let alone the exact range of speech rate changes falling within it. Still, it would be helpful to obtain an overview of how such a gray zone, *if it exists*, would affect the way conversants could actually use the informational potential



(A) Precision (B) Recall

Figure 3.3: Precision and recall of acceleration and deceleration signals with several ranges of gray zones.

of speech rate changes in real dialogues.

For this purpose, we set the range of the gray zone alternatively within ±10 msec, ±20 msec, and ±30 msec, and computed precision and recall rates for each case (see Figure 3.3). The results show that decelerated/accelerated speech is fairly accurate signals for openings/non-openings of information regardless of the range of the gray zone (see Figure 3.3 (A)); while, concerning the recall of deceleration/acceleration signaling, the larger the range of gray zone is, the lower the recall is (Figure 3.3 (B)).

Provided that greater degrees of acceleration/deceleration are perceptually clearer than smaller ones, this finding indicates that we can detect openings and non-openings of information pretty accurately, even when we use only perceptually clear cases of decelerations and accelerations. The comprehensiveness of the detection capability decreases, however, when we rely only on clear cases of speech accelerations and decelerations; this indicate that we will miss out certain cases of information openings and non-openings. Accordingly, there is some trade-off between the perspicuousness of speech rates changes and their comprehensiveness as signals.

This, of course, does not mean that information openings and non-openings are absolutely nondetectable in cases where the degrees of deceleration or ac-

celeration are lower than some perceptibility threshold. In a real conversation, many different events co-occur with accelerations and decelerations, and even when speech rate changes fall in the gray zone, some of those other features may well be used as signals for information structures. For example, prosodic features, such as pitch range, boundary tone, pause, and duration, and linguistic features, such as cue phrases, have been reported to have a relation to discourse structures (Brown et al., 1980; Silverman, 1987; Swerts & Ostendorf, 1997; Hirschberg & Nakatani, 1996; Passonneau & Litman, 1996).

Admittedly, this is but a crude sketch of actual uses of the informational potential of dynamic speech rates by a dialogue conversant, and more exact evaluations would require a combination of results in perceptual acoustics about the range of speech rate changes discernible by average humans, and results in other works involving discourse studies on the informational values of various prosodic and linguistic features of speech. This, however, goes beyond the scope of the present study.

3.4.1.5 Evaluation based on Normalized Average Mora Durations

As we mentioned before, a base AMD might not be the *best* measure of speech rate. Thus, taking the variability in the lengths of morae into account, we re-evaluate the signaling potentials of dynamic speech rates for information structures in terms of precision, recall, and error.

We first calculated the differences between the actual durations of IPUs and the durations of IPUs estimated from the mean length of each phoneme taken from read speech,[8] and then divided the differences by the numbers of morae composing IPUs. We call these values normalized average mora durations (N-AMDs). For example, consider the IPU "mi gi ni," whose actual duration is 416 msec. The mean lengths of phonemes, /m/, /i/, /g/, and /n/ are, respectively, 46 msec, 70 msec, 48 msec, and 41 msec, obtaining 345 msec as the estimated duration of the entire IPU. Dividing the difference between the actual duration and the estimated one, that is, 71 msec, by the numbers of morae, we obtain 23.7

[8]The mean phoneme length data was provided by ATR Interpreting Telecommunications Research Laboratories.

Table 3.10: Frequencies of accelerating/decelerating and opening/non-opening IPU pairs based on normalized average mora durations.

	non-open		open	
Condition	accel	decel	accel	decel
(A)	235	94	49	83
(B)	85	46	13	38
(C)	142	147	52	160
(D)	142	58	–	–
Total	604	345	114	281

as the normalized average mora duration of the IPU.

In the same way as bare AMDs, we compared the N-AMDs of pairs of consecutive IPUs, and classified the pairs of IPUs into accelerating/decelerating. The frequencies of accelerating/decelerating and opening/non-opening IPU pairs are shown in Table 3.10. The associations between accelerating/decelerating and opening/non-opening are all significant for categories (A) to (C). The precision, recall, and error rates are also shown in Table 3.11. The precision and recall rates of acceleration and deceleration are all moderately high, and the error rates are lower than the chance levels (40.9%, 40.3%, and 48.8% for categories (A), (B), and (C), respectively), although the results are not so dramatic as those based on bare AMDs.

Since, as we mentioned before, normalized average mora durations have difficulty in its use for our purpose, including the difference between spontaneous and read speech, the decrease in signaling potentials based on normalized average mora durations is hard to interpret. We, however, believe that the overall tendency obtained here still suggests the significance of signaling potentials of dynamic speech rates to discourse structures.

3.4.2 An Interfering Factor to the Signaling

Apart from the gap between real speech rate changes and perceived speech rate changes, there is another issue to be addressed to obtain a realistic view of the

Table 3.11: Precision, recall, and error rates based on normalized average mora durations.

	Precision of acceleration	Recall of acceleration	Precision of deceleration	Recall of deceleration	Total error
(A)	82.7%	71.4%	46.9%	62.9%	31.0%
(B)	86.7%	64.9%	45.2%	74.5%	32.4%
(C)	73.2%	49.1%	52.1%	75.5%	39.7%
(D)	–	71.0%	–	–	–
Total	84.1%	63.6%	44.9%	71.3%	34.2%

actual communicative function of dynamic speech rates in dialogues. That is, the potential signals by dynamic speech rates are ***neither infallible nor totally comprehensive.***

As Table 3.9 shows, the correlations between dynamic speech rates and information openings/non-openings are only statistical. The acceleration and deceleration signaling allows false alarms of 7.3% and 34.5%, respectively. Consequently, there are cases in which speech rate changes carry *mis*information: a speech acceleration may indicate the absence of an information opening where there actually is, and a deceleration may indicate the presence of an information opening where there is not. In addition, the signaling misses out 18.5% and 15.4% of the cases for acceleration and deceleration, respectively. That is, there are cases in which signals of dynamic speech rates are *absent*: there may be an information opening where there is no speech deceleration that signals it, and some non-openings may be unmarked by speech accelerations.

Table 3.9 shows that the greatest irregularities are to the precision of deceleration signaling in a single speaker's speech: of all type-(B) decelerating pairs, 50.0% are non-opening; of all type-(A) decelerating pairs, 37.9% are non-opening. Let us look at these irregular pairs more closely to find what kinds of extra factors can interfere with the accurate signaling of openings/non-openings of information by dynamic speech rates.

On a closer examination of the immediate contexts of the irregular pairs in

question, we find:

1. 22.9% have their second IPUs preceded by silence of more than 625 msec.[9] The following is an example chosen from our data. (Here, the pairs of IPUs in question are enclosed in boxes. The number in <> placed after an IPU represents the length of silence in milliseconds between the IPU and the succeeding IPU. An English translation of each IPU is shown to its right.)

G: [ima iru tokorokara]	(from where you are now)
F: hai <4256>	(yeah)
G: [nanameueni]	(obliquely above)
F: hai	(yeah)
G: ikimasu	(proceed)

2. 18.1% have their second IPUs followed by silence of more than 625 msec, which in turn is followed by the same speaker's IPU.

G: [sekizoono aidao]	(between the stone statues)
G: [tootte] <1408>	(go through)
G: sekizoo gurutto mawatte kudasai	(go around the stone statue)

3. 34.3% have second IPUs that are exceptionally slow, namely, their AMDs are 180 msec or more.[10]

G: [sosite sokokara-]	(and from there)
G: [ettone (AMD: 320 msec)]	(well)
G: hidari-	(the left)
G: hidarinanameni<576>	(obliquely to the left)

[9] A silence of 625 msec is more significant than it may first appear. The median of the silence duration of our data is 336.0 msec, and 625.0 msec is the third quartile of the frequency distribution of all silences in our data.

[10] The mean AMD of all IPUs in our data is 122.7 msec with a standard deviation of 54.4 msec, and the AMD value of 180 msec is at the +1 standard deviation from it. AMDs of 180 msec or more are "exceptional" in this sense.

4. 25.7% have filled pauses, or fillers, as their second IPUs.

G: hansenno-	(of the sailboat)
G: n-	(uhm)
G: hidarino yonsentikurai	(about four centimeters to the left)

5. 43.8% have some lengthening in or at the end of their second IPUs.

G: e-to koisino kaiganno-	(well, from the pebble beach)
F: hai	(yeah)
G: etto migiue	(well, upper right)
G: zuno sono zuno migiueatarini-	(around the picture's, that picture's upper right)

There are many irregular pairs that have two or more of the structural properties 1–5. The following is a pair of IPUs that has *all five*:

G: syuppatutiten kara-	(from the starting point)
F: hai <864>	(yeah)
G: eetto- (AMD: 331 msec) <1536>	(well)
G: karuku migini hukureruyoona kanzide	(as though you slightly deviate to the right)

A plausible account of these structural properties is that since the cognitive load in producing the second IPU or subsequent IPUs in these cases is heavy, either the speaker delays and slows down the second IPU (case 1), slows down the second IPU and delays the third IPU (case 2), makes the second IPU exceptionally slow (case 3), uses a slowed-down filler before the third IPU (case 4), or lengthens a part of the second IPU (case 5). Overall, 70.5% of the irregular pairs in question have one or more of these structural properties, and could be attributed to the cognitive load in producing the second IPU or its successors.[11]

[11]There are also other irregular pairs whose immediate contexts are none of the above types, but whose global contexts can be considered "stammering speech." Here, we simply characterize the relevant cognitive loads in this and other cases as those in producing second or subsequent IPUs, but this does not preclude the possibility that other kinds of cognitive loads (say, in comprehending a partner's speech, or in memorizing or inferring the configuration of her map) are involved as remote causes.

This suggests the following view on the cuing of information openings by speech decelerations. Generally, in the case of a single speaker's sequential speech, decelerations potentially signal openings of information. This is perhaps also the case of a single speaker's non-sequential speech flanking a voiced backchannel. However, a cognitive constraint on the production of subsequent speech interferes with and slows down the relevant utterance where there is no opening of information. Given the default informational potential of decelerated speech, this slowing down carries misinformation, indicating an opening of information where there is none.

The signaling of information openings and non-openings by speech rate changes is not an isolated, or "shielded," function of speech. There can be various extra constraints in dialogue situations that interfere with the regularities, making the signaling less accurate and comprehensive. For example, in the dialogue situation where a particular issue is immediately settled, the speaker may start presenting the relevant information (or start asking the relevant question) at a greater speed than that of the immediately preceding speech. This would make the acceleration signal less accurate and the deceleration signal less comprehensive. The same regularities are also violated when the speaker "rushes through" a transition relevance place as a strategy to continue his speaking turn (Schegloff 1982, 1987), provided that this particular transition relevance place coincides with the opening of the next information unit. Or more generally, the degree in which a piece of information is preferred or dispreferred at the given point of a dialogue may put certain biases on the speed in which the information is presented. Although we do not have enough space to reexamine our data with respect to these possibilities, the observation in this section suggests limitations on the cognitive resources of speakers as one of the constraints interfering with the precision of deceleration signaling.

3.4.3 Primacy of Dynamic Speech Rate Cuing

3.4.3.1 Spurious Cuing Relation

We demonstrated in the previous sections that local dynamic speech rates of dialogue conversants have the potential of cuing information structures. On the

other hand, there can be also a bundle of other cuing factors, including the local dynamic speech rate, each of which potentially signals information structures of the same type and functions jointly with others to form a reliable signal when all of them are combined. Then, in order to claim that cuing relationship between dynamic speech rate and information structures is a genuine one, we still need to establish that this cuing relation is not a spurious one caused by a more fundamental cuing relationship.

To understand what a spurious cuing relation is, let us see the following examples:

> Meteorologists have found that El Ninõ in the South American coastal sea will bring about, in high probability, both a number of typhoon storms to Japan in the fall, and a wet winter to California. Apparently, therefore, the number of typhoon storms in Japan has the potential of signaling winter weather in California. Clearly, however, the predictive capability of a typhoon storm depends primarily on the occurrence of El Ninõ, and once one knows whether or not El Ninõ will occur, typhoon storms in Japan will no longer be relevant to California weather.

In this case, El Ninõ is the genuine cuing factor for the weather in California, whereas the Japanese typhoon storm is regarded as a spurious cuing factor induced by the relationship between El Ninõ and California weather and that between El Ninõ and Japanese typhoon storm.

3.4.3.2 Verifying Primacy

To demonstrate that dynamic speech rate cuing is not an induced one from a more fundamental cuing, we focus on several features of speech which can have strong relations to information structures. We examine the following two types of features:[12]

[12]There are other cuing factors which can have strong relations to information structures, most notably prosodic features such as pitch range, boundary tone, pause, and duration, which can signal discourse structures of the same type (Brown et al., 1980; Silverman, 1987; Swerts

Temporal features of speech: Those features that are related to and are involved in the computation of dynamic speech rates: (a) absolute speech rate, (b) absolute duration of IPUs, and (c) relative durational changes between successive IPUs.[13]

Lexical features of speech: Choices of lexical categories that make significant differences in the speech rate.

In order to show that the association between dynamic speech rate and information structures is not spurious, we apply the test by M-type elaboration (Yasuda & Umino, 1977). If the association between factor X and factor Y is a spurious one induced by another factor Z, then we observe, in the three-dimensional contingency table regarding the factors X, Y, and Z, the following:

1. There is an association between factors X and Z.
2. There is an association between factors Y and Z.
3. There is no association between factors X and Y in either category of factor Z.

Conversely, if either of the above does not hold, it shows that the association between factors X and Y is not a spurious one induced by factor Z. Thus, focusing on the several rival factors, we examine whether or not the following three conditions are all satisfied.

Condition 1 There is an association between dynamic speech rate and a rival cuing factor.

Condition 2 There is an association between information structures and a rival cuing factor.

& Ostendorf, 1997; Hirschberg & Nakatani, 1996). However, we do not include them in our considerations, because we think they are independent of dynamic speech rate.

[13]The number of morae in IPUs can also be regarded as a candidate factor, but we did not apply our test directly against this factor, because a very high correlation ($r = 0.90$) was observed between IPU durations and the number of morae in IPUs. The test with respect to the IPU durations indirectly apply to the case of the mora number.

Condition 3 There is no association between dynamic speech rate and information structures in either category of a rival cuing factor.[14]

3.4.3.3 Temporal Features of Speech

Absolute Speech Rate: Significant distributional differences were observed in our data among the average speech rates in the starting, intermediate, and ending IPUs; starting IPUs were slower than intermediate IPUs, which in turn were slower than ending IPUs (see Table 3.4). This suggests that an absolute speech rate appears to be a promising candidate for rival cuing factors for information structures. In fact, if the threshold of *slow* IPUs is set to the median (128.8 msec) of the mean AMD of starting IPUs and that of intermediate IPUs, and the threshold of *fast* IPUs is set to the median (109.0 msec) of the mean AMD of intermediate IPUs and that of ending IPUs, then the following two types of cuing relations can be hypothesized:

(a) If an IPU is slow, then it marks an opening; otherwise, it is non-opening.

(b) If an IPU is fast, then it indicates that the next IPU makes an opening; otherwise, the next IPU is non-opening.

We applied the elaboration test to each case of the rival cuing factors (a) and (b). The upper half of Table 3.12 shows the results. From this table, it is found that, in each case, though there is a relatively strong association between dynamic speech rates and the rival factor, there is only a low association between information structures and the rival factor. Moreover, the association between dynamic speech rate and information structures remains strong even when the data is elaborated according to the categories of the rival factor. This result

[14]For example, consider the case of a rival cuing factor being relative durational changes between successive IPUs. We first fix the relative durations either to the "lengthening" category or to the "shortening" category, and then examine whether there is an association between dynamic speech rates and information structures in the fixed category. If there is a strong association in at least one category, then it is regarded as one piece of evidence showing that the association between the dynamic speech rate and information structures is not a spurious one induced by relative durations.

Table 3.12: Results of the elaboration test to temporal rival cuing factors.

Rival cuing factors	Cond. 1	Cond. 2	Cond. 3	
	(ϕ coef.)	(ϕ coef.)	Cuing category	(ϕ coef.)
(a) Absolute speech rate	0.46	0.26	slow	0.51
			not-slow	0.62
(b) Absolute speech rate	0.43	0.34	fast	0.68
			not-fast	0.53
(c) Absolute duration	0.12	0.20	long	0.62
			not-long	0.60
(d) Relative duration	0.26	0.30	lengthening	0.62
			shortening	0.55

means that Conditions 2 and 3 are not satisfied. We can, therefore, say that the association between dynamic speech rate and information structures is not a spurious one induced by the rival cuing factors (a) and (b).

Absolute Durations: The classes of starting, intermediate, and ending IPUs differ not only in their average speech rates, but also in their average durations. The average durations for the three classes of IPUs are 1061.0 msec, 775.7 msec, and 812.6 msec, respectively. A statistical test by ANOVA shows that there is a significant difference of absolute durations among three positions ($F(2, 14) = 8.8$, $p < .01$). Multiple comparison tests (Newman-Keuls procedure) show that starting IPUs are significantly longer than intermediate IPUs ($p < .01$), whereas no significant difference is found between intermediate and ending IPUs. If we set the threshold of *long* IPUs at the median (932.9 msec) of the mean duration of starting IPUs and that of intermediate and ending IPUs, then the following cuing relation is conceivable:

(c) If an IPU is long, then it marks an opening; otherwise, it is non-opening.

The third column in Table 3.12 indicates that none of Conditions 1, 2, and 3 are satisfied. Therefore, we can say that the association between dynamic speech rate and information structures is not an induced one from the rival cuing factor (c).

Relative Durations: Let us call a pair of IPUs *lengthening* if the duration of the second IPU is equal to or greater than that of the first IPU, and call a pair of IPUs *shortening* otherwise. We saw in the previous subsection significant distributional differences of long and short IPUs in starting, intermediate, and ending IPUs. Short IPUs are likely to occupy ending positions, whereas longer IPUs are likely to occupy starting or intermediate positions. Given this difference, the following would be the only cuing relation that can be reasonably hypothesized between relative durations of IPUs and information structures:

(d) If a pair of IPUs is lengthening, then it is an opening pair, and if a pair of IPUs is shortening, then it is a non-opening pair.

The bottom-most column in Table 3.12 indicates that all conditions are not satisfied, and therefore, we can say that the association between dynamic speech rate and information structures is not a spurious one caused by the rival cuing factor (d).

3.4.3.4 Lexical Features

It seems obvious that lexical features of words included in IPUs make a strong cue to the structure of information being presented. Different from temporal features of speech, lexical features are quite different in nature from the local dynamic speech rate, and both factors can make genuine cuing factors coextensive with and comparable in their effectiveness with each other. It is difficult to precisely formulate cuing relations between lexical features and information structures, because of the diversity of lexical features.

Therefore, we first select part of speech categories whose occurrences in IPUs have strong correlations with local accelerations and decelerations of speech, and

set up competing cuing hypotheses in terms of those part of speech categories. We then examine whether or not the three conditions are all satisfied.

Interjections: A series of χ^2 tests singled out interjections, among a set of part of speech categories in Japanese, as a category whose presence or absence brings about significant distributional differences of accelerating and decelerating IPU pairs. We further divide these interjections into two separate categories: interjections that precede main utterances, and interjections that are used in responding to other conversants. The first subcategory of interjections includes fillers and summons: "eeto," "sono," and "ano." The second subcategory includes confirmatory and negative response words and acknowledgments: "hai," "iie," "ee," and "un." We call these two subcategories *precede* type interjections and *response* type interjections, respectively. These two types of interjections display different patterns of correlation with local speech rate changes. If the second IPU contains a precede type interjection, the pair tends to be decelerating (37 vs. 73 for accelerating and decelerating pairs), whereas if the first IPU contains a response type interjection, the pair tends to be decelerating (33 vs. 115 for accelerating and decelerating pairs). Accordingly, two types of interjections work differently in cuing information structures. The following two cuing relations between lexical features of IPUs and the opening and non-opening of information can be hypothesized:

(e) If a precede type interjection appears in an IPU, then it marks an opening; otherwise, it is a non-opening.

(f) If a response type interjection appears in an IPU, then it indicates that the next IPU makes an opening; otherwise, the next IPU is non-opening.

In each case of the rival cuing factors (e) and (f), we examined whether the three conditions are all satisfied. Table 3.13 summarizes the results. From this table, it is found that, in each case, there are only low associations between dynamic speech rates and the rival factor and between information structures and the rival factor. Furthermore, there remain strong associations between dynamic speech rate and information structures in the elaborated situations. This means

Table 3.13: Results of the elaboration test to lexical rival cuing factors.

Rival cuing factors	Cond. 1	Cond. 2	Cond. 3	
	(ϕ coef.)	(ϕ coef.)	Cuing category	(ϕ coef.)
(e) Lexical features	0.17	0.14	precede intj.	0.51
			no precede intj.	0.62
(f) Lexical features	0.29	0.38	response intj.	0.68
			no response intj.	0.55

that none of Conditions 1, 2, and 3 are satisfied. We can, therefore, say that the association between dynamic speech rate and information structures is not an induced one from the rival cuing factors (e) and (f).

Thus, the results of the elaboration test in this section lead to the conclusion that the dynamic speech rate cuing of information structures is not a spurious one caused by other cuing factors.

3.5 Summary

In this chapter, we focused on the global structural elements, discourse structures. In order to elucidate the function of dynamic speech rates as contextualization cues to discourse structures, we analyzed the relation of discourse structures to dynamic speech rates. We examined corpus data of spontaneous dialogues in Japanese, and applied statistical methods to find regular correlations between local changes in the speech rate and the structures of information being expressed. We found correspondences between speech accelerations and the absence of information openings, and between speech decelerations and the presence of information openings. These results show that the dynamic speech rates can potentially function as contextualization cues for openings and non-openings of information expressed in dialogues.

The correlations in question hold not only for a single speaker's utterances but also for multiple speakers' sequential utterances with or without turn changes.

Thus, even when there is a change of speakers during a dialogue, the subsequent speaker decelerates his speech if his speech opens up a new piece of information; otherwise, the subsequent speaker maintains the acceleration pattern established by the preceding speaker. Intentionally or unintentionally, the speakers collaborate with each other to maintain these regularities governing the dynamic speech rate and discourse structures.

Chapter 4

The Relation of Prosodic and Syntactic Features to Turn-Taking and Backchannels

In this chapter, we shift from the global structural elements, and turn to the local structural elements, *turn-taking* and *backchannels*. We first explore the predictive powers of prosodic and syntactic features of the speaker's speech for discriminating turn-taking categories and for discriminating backchannel categories, elucidating the function of these features as contextualization cues to turn-taking and backchannels. We then discuss the interrelationship between prosody and syntax in the decision process of turn-taking and backchannels.

4.1 Introduction

In the previous chapter, we focused on global structural elements, discourse structures. We investigated the relationship between discourse structures and dynamic speech rates in several conditions, showing that changes in the speech rate can potentially function as contextualization cues for the structure of information. In this chapter, we turn to the local structural elements, turn-taking and backchannels, exploring the relation of these phenomena to prosodic and syntactic features of the speaker's speech.

Turn-Taking and backchannels are fundamental phenomena observed in conversational interaction. Numerous studies have been made in various fields such as conversation analysis (Sacks et al., 1974; Goodwin, 1981; Schegloff, 1982), social psychology (Kendon, 1967; Duncan & Fiske, 1977; Beattie, 1983), and discourse analysis (Yngve, 1970; Maynard, 1989; Ford & Thompson, 1996).

One matter of importance in studies on turn-taking and backchannels is how the context of turn-taking and backchannels are characterized by linguistic, paralinguistic, and/or non-linguistic features. Many studies have investigated the relation of turn-taking and backchannels to various features, showing that some features can potentially function as cues for turn-taking and backchannels. For example, some researchers have proposed that changes of turn tend to occur at syntactic or grammatical completion points (Sacks et al., 1974; Duncan & Fiske, 1977; Orestrŏm, 1983; Ford & Thompson, 1996), while turns are likely to be held when grammatical units being constructed are not completed (Ball, 1975). The relation of prosody and body motion to turn-taking has been widely studied as well (Kendon, 1967; Duncan, 1972; Osaka, 1988).

Backchannels have not been studied as much as turn-taking, but a few researchers have suggested that backchannels tend to appear in the context characterized by certain kinds of prosodic and syntactic features of the speaker's speech (Maynard, 1986; Mizutani, 1988; Imaishi, 1994; Ward, 1996); for example, in Japanese conversations, prosodic features such as rising, rise-falling, and flat-falling F0 contours serve as signals for inducing backchannels by the hearer (Imaishi, 1994).

Although many studies have investigated the relation of turn-taking and backchannels to *individual* prosodic and syntactic features, little attention has been given to the *interrelationship* among the features. In particular, the ways in which prosody and syntax contribute to the construction of turns are very complex (Auer, 1996), and therefore, the interrelationship between prosody and syntax in cuing turn-taking and backchannels has opened up various unresolved problems. The investigation of the interrelationship between prosody and syntax in characterizing the context of turn-taking and backchannels would help us to understand the complicated process of turn-taking and backchannels.

In this chapter, we explore prosodic and syntactic features of the speaker's

speech at points where turn-taking and backchannels occur, on the basis of our analysis of Japanese spontaneous dialogues. We first show that prosodic and syntactic features can potentially function as contextualization cues to turn-taking and backchannels. Then, we investigate interrelationship between prosody and syntax in the decision process of turn-taking and backchannels.

This chapter is organized as follows. First, we explain how the corpus data used in this study is analyzed, with a detailed description of the prosodic and syntactic features we make use of. Following the previous studies, we utilize syntactic features involving part of speech, and prosodic features such as duration, patterns of F0 contours and energy trajectories, and peak F0 and energy values. Second, we elucidate the informational potentials of prosodic and syntactic features for turn-taking and backchannels. We examine the relevance of individual features to these phenomena, and then, measure the predictive powers of these features for discriminating turn-taking categories and for discriminating backchannel categories by making use of the decision tree. We show that prosodic and syntactic features can potentially function as contextualization cues to turn-taking and backchannels. Finally, we investigate the interrelationship between prosody and syntax in the decision process of turn-taking and backchannels. We explore the degree of contribution of each feature to discriminating turn-taking categories or backchannel categories by using multivariate statistical analysis techniques, and discuss the interaction between prosody and syntax, comparing them with previous models which have mentioned roles of prosody and syntax in the decision process of turn-taking.

4.2 Methods

4.2.1 Materials

We randomly selected eight dialogues, by sixteen different speakers, from among the Japanese Map Task Corpus. In the Map Task dialogues, two participants exchange utterances spontaneously and naturally, in which they look at similar but significantly different maps of the same region, each unseen by the other. Landmarks on one map may be labeled differently than on the other, or even

lack counterparts. It takes not only the Instruction Giver but also the Follower to take turn very frequently. These dialogues, therefore, can be thought to serve our purpose.

In order to equate the size of each speaker's data, we selected the initial five minutes from each dialogue, where the amount of "five minutes" was determined according to the shortest dialogue in the eight selected ones. In the Map Task, subtasks with the same structures, mainly consisting of identifications of landmarks and instructions to follow paths through the landmarks, tend to be repeated. Therefore, there would be little difference between the initial, intermediate, and final parts of the dialogues, although this itself is an empirical question to be addressed in a future study.

Of the eight dialogues selected for the analysis, four were under the facing condition and the other four under the non-facing condition. From the nature of the task, the subjects usually directed their faces to the maps on the table and seldom to each other, even in the facing condition. It has, in fact, been reported that there are no differences between the facing and non-facing conditions in many aspects of dialogues, such as the time required to finish the task (Nakano et al., 1997). Therefore, although eye-contact might affect turn-taking and backchannels, we do not explore it in this study.

4.2.2 Unit of Analysis

We use an *inter-pausal unit* (IPU) as the unit of analysis (see Section 2.2.1).[1] Table 4.1 provides the basic facts about our data, including the means of the

[1]Intermediate and intonational phrases have often been used as a unit of analysis by researchers in the field of discourse and dialogues (Hirschberg & Nakatani, 1996), although we disprefer them for their difficulty in automatic segmentation. Most of the boundaries of intermediate and intonation phrases are expected to correspond to the boundaries of inter-pausal units. Inter-pausal units, however, might occasionally include the boundaries of intermediate or intonational phrases within them. If turn-taking and backchannels occur at or around these boundaries within IPUs, our analysis based on IPUs will fail to capture such cases. Note, however, that intonational boundaries themselves are not decisive indicators of turn-taking and backchannels. Even if we use them as the unit of analysis, an investigation on the features around the boundaries will be necessary. We will incorporate the use of intermediate and intonational phrases as the unit of analysis in a future study.

Table 4.1: Mean duration of IPUs for each speaker (msec).

	Giver			Follower		
Dialogue	*N*	Mean	SD	*N*	Mean	SD
1	175	861.7	607.5	120	691.6	763.9
2	134	991.1	636.6	86	524.0	410.0
3	95	1160.4	727.7	100	778.5	762.4
4	149	822.1	559.9	91	770.3	530.5
5	119	970.4	681.2	116	700.8	631.3
6	145	995.2	720.9	126	468.8	417.5
7	176	813.6	651.6	85	522.3	470.6
8	136	1017.6	671.1	102	772.8	702.3

duration of IPUs in each dialogue.

4.2.3 Turn Transition Types

We classify pairs of consecutive IPUs into two turn-taking categories, namely, CHANGE and HOLD, according to whether the hearer or the speaker of the first IPU produces the second one. We further classify the HOLD type pairs into two sub-types, namely, BC and NO-BC, according to whether or not they are accompanied by backchannels (see Figure 4.1).[2]

We left out cases in which two conversants spoke concurrently. That is, we did not give full analysis to cases of two "overlapping" inter-pausal units, because our focus in this chapter is placed on *smooth* turn transitions. It would, however, be an interesting extension of this study to investigate the relationship between

[2]A backchannel is a short utterance, such as "yeah" and "uh huh" in English, uttered by the hearer without claiming the shift of turn (Maynard, 1989; Schegloff, 1982). We judged backchannels from their forms and functions. Interjectory expressions such as "hai," "ee," and "un" in Japanese were judged to be backchannels unless they constituted conversational moves such as an answer to a yes-no question (Carletta et al., 1997) (see Noguchi (1998) for more concrete definition of backchannels). Two labelers judged backchannels based on the above criteria, and cross-checked each other's results.

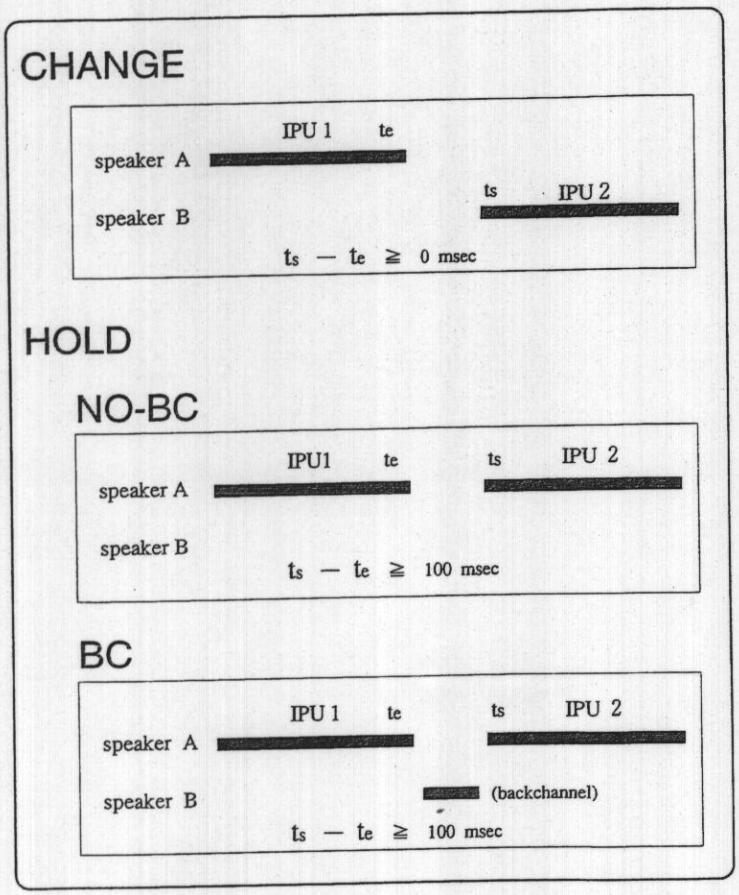


Figure 4.1: Turn transition types.

non-smooth turn transitions and the features of the speaker's speech in order to see the difference between smooth and non-smooth transitions.[3]

4.2.4 Prosodic and Syntactic Features

In this section, we explain the prosodic and syntactic features of Japanese utterances we are concerned with. In this study, we assume that the information relevant to turn-taking and backchannels is localized at a point just before a turn change or a backchannel, that is, at the end of IPUs. We label each IPU with six features derived from the properties of the final portion of the IPU. The six features are:

1. the part of speech of the final morpheme,
2. the duration of the final phoneme,
3. the pattern of the F0 contour in the final mora region,
4. the relative height of the peak F0 of the final mora,
5. the pattern of the energy trajectory in the final phoneme region, and
6. the relative height of the peak energy of the final phoneme.

In Japanese, intonation patterns, at sentence or clause boundaries, are relevant to modality and discourse functions (Kôri, 1997; Inoue, 1997). For example, falling and rising patterns are relevant not only to the ending of sentences but also to the declarative and interrogative modes. Rise-falling patterns, on the other hand, express non-finality of sentences (Kôri, 1997), sometimes with the function of pushing the dialogue ahead. Intonation patterns have also been reported to relate with turn-taking and backchannels (Inoue, 1997; Imaishi, 1994). These intonation patterns are mainly connected with F0 contours, but energy and durational aspects might also be relevant. We therefore include all these factors.

[3] An analysis of non-smooth transitions from a somewhat different viewpoint will be presented in Chapter 5.

In regard to the F0 features, we focus on the final mora region, because it is said that, in Japanese, intonation patterns can be attributed to an F0 contour in the sentence, clause, or phrase final region, which corresponds approximately to the final mora (The National Language Research Institute, 1960). Prolongation at boundaries of syntactic phrases or clauses and changes in the loudness, on the other hand, are mainly associated with vowels, which suggests the significance of the region of the final phoneme in regard to the duration and energy features.

In this study, the features are not represented as continuous values but as categorical ones, because some features are difficult to describe as continuous values. For example, rise-falling F0 patterns cannot be represented only by onset or offset F0 values. The use of categorical values also make the analysis on the interrelationship among the features easy and precise, because parts of speech are inherently categorical and it would be difficult to handle categorical and continuous values together.

Below, we explain the possible values of the features and the way they are assigned.

4.2.4.1 Parts of Speech

The final morpheme of each IPU is tagged with one of the seventeen parts of speech shown in Table 4.2.

The leftmost column in Table 4.2 includes non-inflectional categories. Nouns and adverbs are as usual. Adnouns modify nouns, e.g., "aru (some)" in "aru hito (someone)." Conjunctions function as the connection between two phrases, clauses, or sentences. Interjections include confirmatory and negative responses and acknowledgments such as "yes" and "no" in English, sometimes functioning as backchannels. As for filled pauses, or fillers, we can list words or expressions such as "eeto" and "ano" in Japanese. Since filled pauses have been pointed out to have particular functions for turn-taking (Brown, 1977), we classify them as a separate category.

The inflectional categories, shown in the middle column in Table 4.2, include verbs, adjectives, and auxiliary verbs, which inflect in seven forms. Verbs in conclusive and imperative forms usually appear at the end of sentences, and

Table 4.2: Part of speech features.

Non-inflectional	Inflectional	Particles
noun	v-irr (irrealis form)	p-case (case/adverbial)
adverb	v-adv (adverbial form)	p-conj (conjunctive)
adnoun	v-euph (euphonic form)	p-final (sentence-final)
conjunction	v-conc (conclusive form)	p-intj (interjectional)
interjection	v-attr (attributive form)	
filler	v-cond (conditional form)	
	v-imp (imperative form)	

constitute declarative and imperative sentences, respectively. Verbs in attributive form modify nouns, e.g., "utau (singing)" in "utau syoozyo (a singing girl)," and verbs in conditional form, followed by particles, constitute hypothetical clauses. Verbs in irrealis form occur at the mid-sentential position and are followed by auxiliary verbs expressing negation, passivization, causativization, and so on. Verbs in adverbial form constitute phrases, clauses and sentences, with or without auxiliary verbs, whereas those in euphonic form are always followed by auxiliary verbs (past tense and politeness) or conjunctive particles.

Postpositional particles (the rightmost column in Table 4.2) are classified into four sub-categories: (i) case and adverbial particles, including topic particles, (ii) conjunctive particles, (iii) sentence-final particles, and (iv) interjectional particles. Case/adverbial, conjunctive, and sentence-final particles usually mark the end of phrases, of clauses, and of sentences, respectively. Interjectional particles occur at the end of phrases and clauses, preceded by other types of particles if they exist, and have the function of calling for the hearer's attention.

4.2.4.2 Durations, F0, and Energy

The values of the prosodic features were assigned as follows. First, the fundamental frequency, energy, and spectrogram of each IPU were extracted at 8 msec intervals using the Entropic Systems ESPS/Waves+ software. A median filter was

applied to the F0 data in the voiced region, and this removed much of the perturbation caused by laryngealization. The energy data was not smoothed. Next, referring to the spectrograms, F0 contours, and energy trajectories, two labelers identified the boundaries of the final mora and the final phoneme of the IPU, and cross-checked each other's results.[4] Finally, based on these boundaries, the values of the five prosodic features were estimated automatically in the following way:

Duration The mean duration of the final phoneme over all IPUs was calculated,[5] and thresholds of 'mean $\pm$ 1 s.d.' were used to classify phoneme durations into "short," "normal," or "long" types.

F0 pattern First, we approximated F0 contours in the final mora region by connecting two approximating lines using the least squares method, and obtained their slopes and their intersection point. Next, each of the two lines was classified into "rising," "falling," or "flat" automatically according to the thresholds which were determined by the two labelers' intuitions.[6] Finally, the overall F0 patterns were classified into "rising," "falling," or "flat" according to the type of the line occupying more than half of the region, with the exception of "rise-fall" and "flat-fall," which were used when the second line was "falling" and the first line of "rise" or "flat" type occupied more than a quarter of the region (the two types were distinguished according to the type of the first line).

Peak F0 We classified the peak F0 values of the final morae into "high" or "low" according to whether the peak was above or below the center of the F0 range of the speaker.

Energy pattern Like F0 patterns, we approximated energy trajectories in the final phoneme region by connecting two approximating lines, obtained their slopes and their intersection point, and automatically classified each line

[4]We exclude those IPUs whose final morae are devoiced.

[5]Logarithmic transformation was employed to satisfy the normality of the distribution.

[6]Thresholds for rising and falling F0 contours should be more precisely determined based on perception experiments over many subjects, but this is beyond the scope of the present study.

into "decreasing" or "non-decreasing" according to the thresholds determined by the two labelers' intuitions. The overall energy patterns were classified into "decreasing," "late-decreasing," or "non-decreasing" types by the following rules. If the first line is of the "decreasing" type, then the overall type is "decreasing"; if both lines are of "non-decreasing," then the overall type is "non-decreasing"; otherwise, the overall type is one of the three types according to the position of the intersection point, i.e., "decreasing" if it is within the first one-third of the phoneme region, "late-decreasing" if within the second, or "non-decreasing" if within the last.

Peak energy We classified the peak energy values of the final phoneme into "high" or "low" according to whether the peak was above or below the center of the energy range of the speaker.

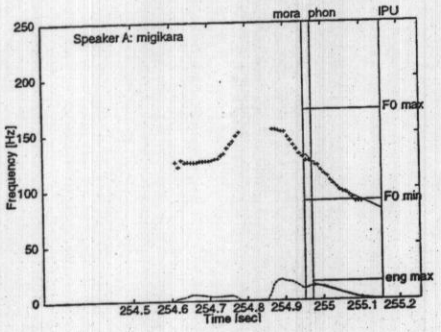
Figure 4.2 shows samples of F0 and energy patterns with the approximating lines. They also indicate the boundaries of the final mora and phoneme, the top line and the baseline of F0, and the top line of energy of the speaker.

As the result of the labeling, we obtained 379 IPUs for CHANGE type, and 654 for HOLD type (121 and 533 for BC type and NO-BC type, respectively), excluding cases whose features could not be reliably determined.

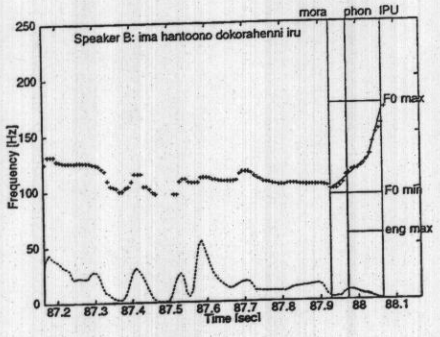
4.3 Signaling Potentials for Turn-Taking

In this and the following sections, we elucidate the informational potentials of prosodic and syntactic features as contextualization cues to turn-taking and backchannels. We first analyze the prosodic and syntactic features appearing at the end of the IPUs in terms of their correlations to turn transition types and to the presence/absence of backchannels, and then, measure the predictive powers of these features for discriminating turn-taking categories and for discriminating backchannel categories by making use of the decision tree.

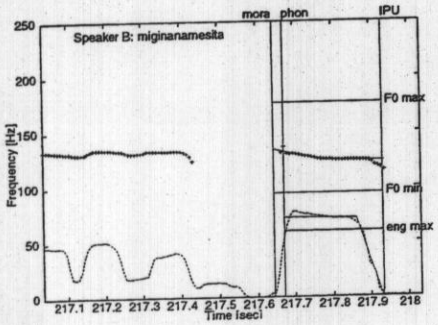
Our concern in this section is turn-taking, that is, the relation of the features to the two turn transition types, CHANGE and HOLD, the latter including both BC and NO-BC (see Figure 4.1). We first examine the relation of individual features to turn-taking, and then, measure the predictive power of these features.



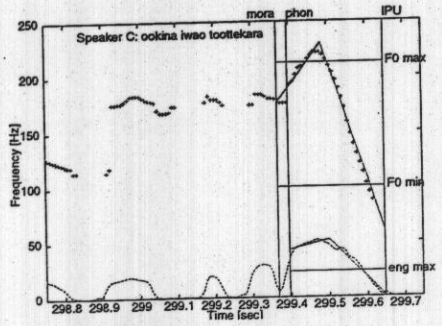
(a) F0 pattern: falling, Peak F0: low, Energy pattern: decreasing, Peak energy: high, Duration: normal



(b) F0 pattern: rising, Peak F0: high, Energy pattern: decreasing, Peak energy: low, Duration: normal



(c) F0 pattern: flat, Peak F0: low, Energy pattern: non-decreasing, Peak energy: high, Duration: long



(d) F0 pattern: rise-fall, Peak F0: high, Energy pattern: late-decreasing, Peak energy: high, Duration: long

Figure 4.2: Sample F0 and energy patterns. The upper curves represent F0 contours and the lower ones energy trajectories. The straight lines on the curves approximate the F0 contours and the energy trajectories. The vertical lines labeled as "mora," "phon," and "IPU" indicate the boundaries of the final mora, the final phoneme, and the IPU, respectively. The horizontal lines labeled as "F0 max," "F0 min," and "eng max" indicate the top line and the baseline of F0 and the top line of energy of the speaker, respectively.

4.3.1 Relation of Syntactic Features to Turn-Taking

The relationship between turn-taking and prosodic and/or syntactic features have never been fully investigated in Japanese conversations. Here, we start with analyzing syntactic features and their relations to turn-taking.

We first calculated the frequencies of the syntactic features relative to the two turn-taking categories, CHANGE and HOLD. For each feature, we then tested the significance of the difference of the frequencies between the two categories using a binomial test, with the ratio in population being 379:654, that is the ratio of the overall frequencies of the two types. Table 4.3 shows the results. It was found that some syntactic features strongly coincide with CHANGE and others with HOLD. Excluding the cases with low frequencies (less than 30), 80% (8 out of 10) of the features were related to either of the categories. Therefore, we can conclude that the syntactic features being examined are one of the major factors relevant to turn-taking.

Let us see the results more closely. In the previous studies, it has been proposed that turn changes tend to occur at syntactic or grammatical completion points (Duncan & Fiske, 1977; Orestrom, 1983; Ford & Thompson, 1996; Maynard, 1989), while they tend not to occur at grammatical incompletion points (Ball, 1975). In Table 4.3, the features associated with CHANGE are, without exception, "sentence-final elements," i.e., verbs in imperative and conclusive forms, sentence-final particles, and interjections, where a "sentence-final" can be defined as the location at which the sentence terminates in, at least, written Japanese. Since we are investigating only *smooth* transitions, this result might not be so surprising; speakers would supposedly complete their turns in such a way.

Associated with HOLD, on the other hand, are conjunctions, adverbs, filled pauses, and case particles, which can be called "mid-sentential elements." In particular, filled pauses have been considered to function as turn-holders (Ball, 1975; Brown, 1977). Our results are in agreement with those studies. There are also some "neutral" features, namely, nouns and conjunctive particles, which coincide with neither CHANGE nor HOLD. These elements are not *typical* sentence-final elements, at least, in written Japanese, but can appear at sentence, or utterance, final positions in spoken Japanese, when accompanied by falling or

Table 4.3: Frequencies of syntactic features relative to turn-taking categories and results of binomial tests.

Feature	CHANGE	HOLD	Binomial test
intj	**107**	61	CHANGE > HOLD**
v-conc	**57**	6	CHANGE > HOLD**
v-imp	**15**	3	CHANGE > HOLD**
p-final	**71**	6	CHANGE > HOLD**
noun	46	75	n.s.
adnoun	0	11	n.s.
v-adv	10	12	n.s.
v-euph	0	1	n.s.
v-attr	0	3	n.s.
v-cond	2	17	n.s.
p-conj	27	59	n.s.
p-intj	3	18	n.s.
adverb	3	**46**	CHANGE < HOLD**
conj	2	**43**	CHANGE < HOLD**
filler	2	**57**	CHANGE < HOLD**
p-case	35	**231**	CHANGE < HOLD**
Total	379	654	

(**: $p < .01$, n.s.: not significant at 1% level)

rising intonation contours.

4.3.2 Relation of Prosodic Features to Turn-Taking

The relationship between turn-taking and the prosodic features was analyzed in the same way. Table 4.4 shows the frequencies of the prosodic features relative to the two turn-taking categories as well as the results of binomial tests. In parallel to the syntactic features, there are two groups of the prosodic features coincident with CHANGE and HOLD, respectively. Many prosodic features belong to either of the two groups. We can, therefore, conclude that almost all of the prosodic features being examined are relevant to turn-taking.[7]

Some researchers have suggested that the rising and falling F0 patterns are related to changes of turn (Duncan & Fiske, 1977; Ford & Thompson, 1996), while the rise-fall and flat-fall F0 patterns tend to be related to continuations of turn (Kôri, 1997; Inoue, 1997). Table 4.4 shows that our results support their findings. Kôri (1996) suggested that, in Japanese, prolonged morae at the boundaries of syntactic clauses also indicate the continuation of the same speaker's turn. This is consistent with our finding that HOLD frequently occurs when the final phoneme of an IPU has a long duration. As Osaka (1988) showed, energy is also an important factor relevant to turn-taking; low or decreasing energy is associated with CHANGE, and high, late-decreasing, or non-decreasing energy is associated with HOLD.

[7]As mentioned before, it is difficult to represent F0 and energy patterns as continuous values. However, the other prosodic features, i.e., duration, peak F0, and peak energy, can be described as continuous values, and can thus be more precisely examined by using a t-test. The mean durations of final phonemes of IPUs for CHANGE and HOLD were 142 msec and 226 msec respectively, confirming a significant difference ($t(1031) = 11.05$, $p < .01$). The means of the peak F0 and peak energy, normalized between 0 and 1 according to the speaker's F0 and energy ranges, also confirmed significant differences for CHANGE and HOLD (peak F0: 0.36 vs. 0.60, $t(1031) = 7.16, p < .01$; peak energy: 0.48 vs. 0.77, $t(1031) = 10.54, p < .01$). These results are in agreement with those shown in Table 4.4. In the subsequent analyses, we will use categorical values for these features for the ease in analyzing the interrelationship among features.

Table 4.4: Frequencies of prosodic features relative to turn-taking categories and results of binomial tests.

Feature	CHANGE	HOLD	Binomial test
Duration			
short	**90**	50	CHANGE > HOLD**
normal	268	378	n.s.
long	21	**226**	CHANGE < HOLD**
F0 pattern			
fall	**184**	126	CHANGE > HOLD**
rise	**68**	14	CHANGE > HOLD**
flat	95	**292**	CHANGE < HOLD**
flat-fall	24	**162**	CHANGE < HOLD**
rise-fall	8	**60**	CHANGE < HOLD**
Peak F0			
low	276	389	n.s.
high	103	**265**	CHANGE < HOLD**
Energy pattern			
decr	**297**	337	CHANGE > HOLD**
late-decr	64	**245**	CHANGE < HOLD**
non-decr	18	**72**	CHANGE < HOLD**
Peak energy			
low	**238**	215	CHANGE > HOLD**
high	141	**439**	CHANGE < HOLD**
Total	379	654	

(**: $p < .01$, n.s.: not significant at 1% level)

4.3.3 Evaluation of the Signaling Potentials

In the previous sections, we examined the relation of individual prosodic and syntactic features to turn-taking, showing that these features were all related to turn-taking, and that the way they correlated was fairly consistent with previous studies. These results give us insight into the kinds of properties prosodic and syntactic features have with respect to turn-taking. In this section, we measure how accurately these features in combination predict the two turn transition types, namely, CHANGE and HOLD.

As a measurement of the predictive power, we make use of the error rates of decision tree. We constructed a decision tree for discriminating CHANGE and HOLD, using C4.5 decision tree program (Quinlan, 1992), based on the six features, and evaluated its performance with eight-fold cross-validation where the data from each dialogue was used, in turn, as the test data and the remaining data was used for training. The left column in Table 4.5 shows the results. The table shows the error rates for the pruning certainty factor (CF) value with which the best performance was achieved. It also shows the error rate for the inside data, which is the performance when the test data is also used for training. The error rate is 16.7% for the inside data and 21.2% for the cross-validation, which are much better than the chance level, 46.5%.

Since the decision tree program we are using can deal with continuous features as well as categorical ones, we can also examine the case where some of the prosodic features, i.e., duration, peak F0, and peak energy, are represented as continuous values. The right column in Table 4.5 shows the results. We obtained 12.7% for the inside data and 21.7% for the cross-validation, which are not so different from the results for categorical variables only.

These results show that the prosodic and syntactic features have strong predictive power for discriminating CHANGE and HOLD, no matter whether the features are categorical or continuous. We can, therefore, conclude that the prosodic and syntactic features can function as accurate cues to turn-taking.

Table 4.5: Error rates of the decision trees predicting the turn-taking categories. (chance level of the error rate = 46.5%)

	Without continuous variables (CF value = 30%)	With continuous variables (CF value = 20%)
cross-validation	21.1%	21.7%
inside data	16.7%	12.7%

4.4 Signaling Potentials for Backchannels

Let us, now, shift our concern from turn-taking to backchannels. In this section, we analyze the relation of prosodic and syntactic features of the speaker's speech to the production of backchannels by the hearer. We consider the two turn transition types, the presence of backchannels, BC, and the absence of backchannels, NO-BC, which have been treated as a single category, i.e., HOLD, in the previous section (see Figure 4.1). In the same way as turn-taking, we examine, first, the relation of individual features to backchannels, and then, measure the predictive power of these features for discriminating the presence and the absence of backchannels by making use of the decision tree.

4.4.1 Relation of Syntactic Features to Backchannels

The relationship between backchannels and the syntactic features was analyzed in the same way as turn-taking. Table 4.6 shows the frequencies relative to the two backchannel categories and the results of binomial tests.

Some researchers have proposed that the presence of backchannels is associated with certain kinds of syntactic features; for example, Mizutani (1988) reported that conjunctive particles are associated with the presence of backchannels. Also Imaishi (1994) made a similar claim but extended her observation to cover other cues to backchannels, such as interjectional and case particles, as well. Our results are in agreement with their findings. We found that conjunctive

Table 4.6: Frequencies of syntactic features relative to backchannel categories and results of binomial tests.

Feature	BC	NO-BC	Binomial test
v-adv	**7**	5	BC > NO-BC**
p-case	**59**	172	BC > NO-BC**
p-conj	**28**	31	BC > NO-BC**
noun	16	59	n.s.
adnoun	0	11	n.s.
v-euph	0	1	n.s.
v-conc	0	6	n.s.
v-attr	0	3	n.s.
v-cond	6	11	n.s.
v-imp	0	3	n.s.
p-final	0	6	n.s.
p-intj	5	13	n.s.
adverb	0	**46**	BC < NO-BC**
conj	0	**48**	BC < NO-BC**
intj	0	**61**	BC < NO-BC**
filler	0	**57**	BC < NO-BC**
Total	121	533	

(**: $p < .01$, n.s.: not significant at 1% level)

and case particles, as well as verbs in adverbial form, tend to coincide with the appearance of backchannels. However, our study goes beyond theirs in that we did not only study the context for the presence of backchannels but also for the *absence* of backchannels. The previous studies have not fully discussed the context for absence of backchannels, except for few mentions about the role of filled pauses in inhibiting backchannels (Sugitô, 1989). Our analysis suggests that the production of backchannels by the hearer is suppressed by such syntactic features as adverbs, conjunctions, interjections, and filled pauses.

4.4.2 Relation of Prosodic Features to Backchannels

Table 4.7 shows the frequencies of the prosodic features relative to the two backchannel categories and the results of binomial tests.

It has been argued that the appearance of backchannels is associated with several kinds of prosodic features, as well as syntactic features; for example, Imaishi (1994) and Inoue (1997) found that flat-fall and rise-fall F0 patterns are likely to be followed by backchannels. Our analysis completely agrees with these observations. Furthermore, Imaishi (1994) claimed that falling F0 patterns are sometimes followed by backchannels but are rather weak indicators, while flat F0 patterns never coincide with backchannels. These observations were also reproduced.

Little attention has been given to prosodic features other than the fundamental frequency. Our study advances the situation by taking several other prosodic features into account. Table 4.7 shows the relations of low or non-decreasing energy to the absence of backchannels.

4.4.3 Evaluation of the Signaling Potentials

In the previous sections, we examined the relation of individual prosodic and syntactic features to the presence/absence of backchannels, showing that, many, though not all, of the features were related to backchannels. In this section, we measure how accurately these features can serve as signals to discriminate the presence and the absence of backchannels, namely BC and NO-BC.

Table 4.7: Frequencies of prosodic features relative to backchannel categories and results of binomial tests.

Feature	BC	NO-BC	Binomial test
Duration			
short	3	47	n.s.
normal	68	310	n.s.
long	50	176	n.s.
F0 pattern			
flat-fall	**48**	114	BC > NO-BC**
rise-fall	**26**	34	BC > NO-BC**
rise	4	10	n.s.
fall	11	**115**	BC < NO-BC**
flat	32	**260**	BC < NO-BC**
Peak F0			
low	80	309	n.s.
high	41	224	n.s.
Energy pattern			
decr	57	280	n.s.
late-decr	60	185	n.s.
non-decr	4	**68**	BC < NO-BC**
Peak energy			
high	100	339	n.s.
low	21	**194**	BC < NO-BC**
Total	121	533	

(**: $p < .01$, n.s.: not significant at 1% level)

Table 4.8: Error rates of the decision trees predicting the backchannel categories. (chance level of the error rate = 30.2%)

	Without continuous variables (CF value = 90%)	With continuous variables (CF value = 65%)
cross-validation	21.5%	22.6%
inside data	12.8%	9.8%

In the same way as turn-taking, we constructed a decision tree for discriminating BC and NO-BC based on the six features, and evaluated its performance with eight-fold cross-validation. The left column in Table 4.8 shows the results. The error rate is 12.8% for the inside data and 21.5% for the cross-validation, which are better than the chance level, 30.2%.

The right column in Table 4.8 shows the results in the case where some prosodic features are represented as continuous values. We obtained 9.8% for the inside data and 22.6% for the cross-validation, which are comparable to the results for categorical variables only.

These results show that the prosodic and syntactic features have strong predictive power for discriminating BC and NO-BC. We can, therefore, conclude that the prosodic and syntactic features can potentially function as contextualization cues to backchannels.

4.5 Interrelationship between Prosody and Syntax

In the previous sections, we explored the predictive power of prosodic and syntactic features for discriminating turn-taking categories or backchannel categories, elucidating the function of these features as contextualization cues to turn-taking and backchannels. In this section, we shift emphasis away from the signaling potentials of these features, and turn to the interrelationship between prosody

and syntax in the decision process of turn-taking and backchannels. Such investigation into the interrelationship among the features has never been addressed. We first investigate how strongly each feature contributes to discriminating turn transition types and the presence/absence of backchannels, and then, discuss the interaction between prosody and syntax in the decision process of turn-taking and backchannels.

4.5.1 Turn-Taking

4.5.1.1 Contribution of Features to Discriminating Turn-Taking Categories

To measure and compare the degree of contribution of each feature to discriminating turn-taking categories, we make use of two sorts of multivariate statistical analysis techniques: one is Quantification II (Hayashi, Diday, Jambu, & Ôsumi, 1988) and the other is the feature deletion test by the decision tree (Haruno, Shirai, & Ôyama, 1998). The former is a variant of discriminant analysis for categorical data, which provides us the measurement for the degree of the contribution of *each feature value*; the latter provides the measurement for the degree of the contribution of *each feature*.

First, we explore how strongly each feature value contributes to discriminating turn transition types making use of Quantification II. Table 4.9 summarizes the results. A large positive value of the normalized category score indicates a strong contribution of that feature to CHANGE, whereas a large negative value indicates a strong contribution to HOLD. From the table, we can see the following: (1) syntactic features such as verbs in imperative and conclusive forms and sentence-final particles have extremely strong contributions to CHANGE, (2) syntactic features such as adverbs and verbs in attributive form have strong contributions to HOLD, (3) syntactic features such as nouns and conjunctive particles have virtually no contributions to either CHANGE or HOLD, (4) rising F0 pattern has moderate contributions to CHANGE, and (5) the other prosodic features have weak or no contributions. Therefore, we can say that some instances of syntactic features make extremely strong contributions to discriminating CHANGE and HOLD, compared with prosodic features other than rising F0 patterns.

Table 4.9: Results of Quantification II: normalized category scores assigned to syntactic and prosodic features with regard to the discrimination of turn-taking categories (correlation rate = .45, discrimination rate = 80.7%).

Feature	Normalized category score
Part of speech = p-final	1.37
Part of speech = v-conc	1.24
Part of speech = v-imp	1.04
F0 pattern = rise	0.64
Part of speech = intj	0.49
F0 pattern = fall	0.33
Part of speech = v-adv	0.28
Duration = short	0.24
Peak F0 = low	0.08
Energy pattern = decr	0.05
Peak energy = low	0.05
Duration = normal	0.01
Part of speech = noun	−0.02
Part of speech = p-conj	−0.02
Peak energy = high	−0.04
Energy pattern = late-decr	−0.04
Peak F0 = high	−0.14
Duration = long	−0.17
F0 pattern = flat	−0.20
Energy pattern = non-decr	−0.24
F0 pattern = rise-fall	−0.29
F0 pattern = flat-fall	−0.33
Part of speech = p-case	−0.46
Part of speech = v-cond	−0.50
Part of speech = p-intj	−0.57
Part of speech = filler	−0.70
Part of speech = conj	−0.83
Part of speech = adnoun	−0.89
Part of speech = adverb	−0.93
Part of speech = v-attr	−1.03
Part of speech = v-euph	−1.59

This, however, does not mean predominance of syntax over prosody, nor even over any of the individual prosodic features. Quantification II can only compare contributions of *feature values*, e.g., a comparison between "part of speech = noun" and "F0 pattern = falling," but it can say nothing about comparisons among *features*, e.g, a comparison between "part of speech" and "F0 pattern."

Therefore, to examine the degree of contribution of *features*, not feature values, we next compare the predictive powers of sets of features, each set constructed by removing one particular feature from the whole set of features (Haruno et al., 1998). We use the error rate of the decision tree derived from a set of features as a measurement of the predictive power.[8]

We, first, obtained the error rate of the decision tree constructed from all six features being analyzed. This gives us the baseline error rate. For each feature X, we then obtained the error rate of the decision tree constructed from the five features other than X. Finally, we calculated the increase in the error rate of the X-removed tree compared with the baseline. A large increase indicates a large contribution of feature X.

The left column in Table 4.10 shows the results. It is found that the increase in the error rate for the syntactic feature (6.3%) is much larger than that for any of the prosodic features (1.1%, 2.5%, 0.5%, 1.3%, and 0.7%), suggesting that syntax has a stronger contribution to discriminating turn-taking categories than any individual prosodic feature.

Concerning such a difference between prosody and syntax in their relative contributions to discriminating turn transition types, one question still remains, that is, whether, and how much, prosody *as a whole* contributes to discriminating turn transition types. To answer this question, we obtained the error rate of the decision tree constructed from the syntactic features only, excluding all of the prosodic features. This allowed us to measures the contribution of prosody as a whole. The bottom-most row in Table 4.10 shows the result. The increase in the

[8]The aim of the use of the decision tree in the previous sections was to elucidate the signaling potentials, so we provided evaluation results for cross-validation. However, the aim here is not such a purpose of evaluating the performance but to extend our knowledge of dependencies among the features. For this reason, we do not address any evaluation issues in this section, and simply rely on error rates for the inside data.

Table 4.10: Error rates of X-removed decision trees predicting the turn-taking categories for each feature X. The increase in the error rate compared to the baseline is shown in parentheses.

	Without continuous variables	With continuous variables
Baseline	14.4%	10.2%
Part of speech	20.7% (6.3%)	17.3% (7.1%)
Duration	15.5% (1.1%)	12.4% (2.2%)
F0 pattern	16.9% (2.5%)	14.1% (3.9%)
Peak F0	14.9% (0.5%)	12.3% (2.1%)
Energy pattern	15.7% (1.3%)	14.5% (4.3%)
Peak energy	15.1% (0.7%)	12.6% (2.4%)
Whole prosody	19.8% (5.4%)	19.8% (9.6%)

error rate for all prosodic features (5.4%) is comparable to that for the syntactic feature (6.3%).

We can also examine the case where some of the prosodic features are represented as continuous values. The right column in Table 4.10 shows these results. The basic observation about the predominance of syntax over any individual prosodic feature is reliably replicated, although prosody as a whole, in this case, has, in some degree, a stronger contribution to turn-taking than syntax. (The increase in the error rate for syntax is 7.1%, while that for prosody is 9.6%.)

From these results, we can say that, in turn-taking, (1) some instances of syntactic features have extremely strong contributions, and (2) in general, syntax has a stronger contribution than any individual prosodic feature, although the whole prosody contributes as strongly as, or even more strongly than, syntax.

4.5.1.2 Schema for Discriminating Turn-Taking Categories

Based on the results presented so far, we investigate the way in which prosody and syntax are related to turn-taking.

Some researchers have discussed the relationship between prosody and syntax

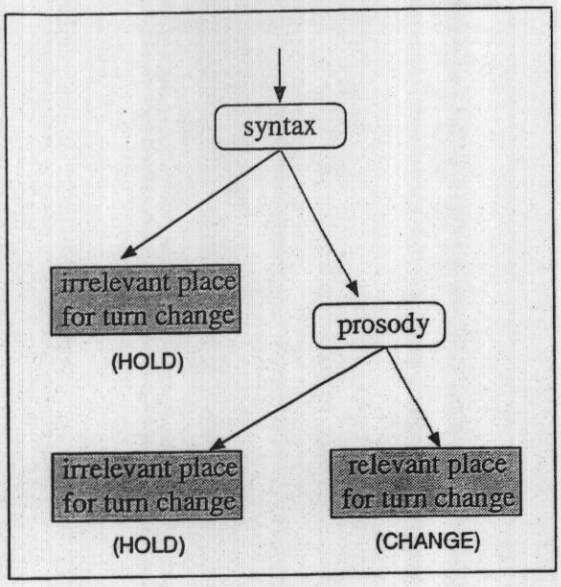


Figure 4.3: Filter model.

with respect to turn-taking. One of the most influential views on this is what Auer (1996) calls the "filter model," where prosody is thought to narrow down the candidates selected by syntax as a relevant place for changes of turn (Schegloff, 1996; Ford & Thompson, 1996).[9] Hence, prosody can be regarded as a "filter" between syntax and turn-taking. This model can be schematized as Figure 4.3 shows.

In order to discuss the relationship between prosody and syntax in our data in comparison with the filter model, we construct a similar schema based on our results. Let us recall the results of Quantification II (Table 4.9). Syntactic features such as sentence-final particles and verbs in imperative and conclusive forms are extremely strong discriminators of CHANGE, while features such as adverbs and verbs in attributive form are extremely strong discriminators of HOLD, suggesting that these features of IPUs might be able to discriminate whether a turn is

[9]Auer (1996) himself, however, claimed the relation of prosody to syntax is very complex and the filter model is not adequate.

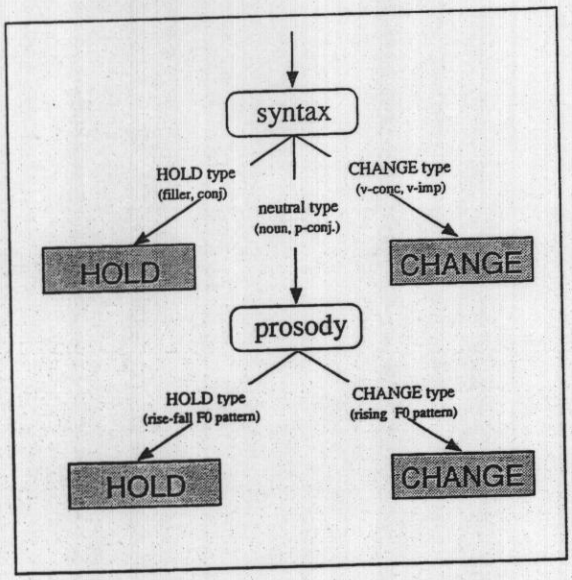


Figure 4.4: Schema for the decision process of turn-taking.

changed or held, no matter which prosodic features co-occur. On the other hand, syntactic features such as nouns and conjunctive particles do not have strong contributions to either CHANGE or HOLD, and hence we can suppose that in these cases, prosodic features as well come into play in deciding the turn transition type.

These results can be summarized as in the following schema for the decision process of turn transition types (see Figure 4.4). (i) If a syntactic feature strongly connected with CHANGE appears, then the turn tends to be changed; (ii) if a syntactic feature strongly connected with HOLD appears, then the turn tends to be held; (iii) otherwise, (a) if a prosodic feature strongly connected with CHANGE appears, then the turn tends to be changed, and (b) if a prosodic feature strongly connected with HOLD appears, then the turn tends to be held.

This schema can be supported by the pattern of the decision tree constructed from all of the six features being analyzed (Figure 4.5). Some of the CHANGE cases can be determined on the basis of syntactic features only. The same can be pointed out to the case of HOLD. However, there are also cases where turn

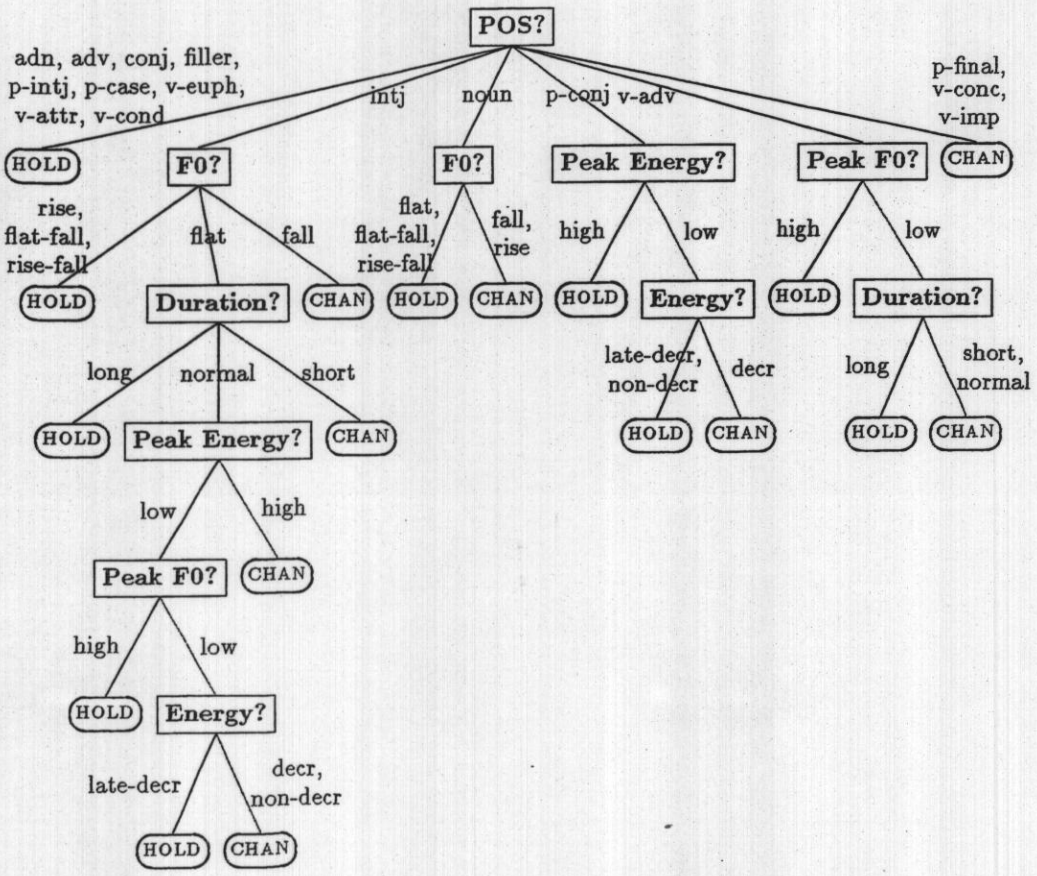


Figure 4.5: A decision tree predicting turn-taking categories from syntactic and prosodic features (error rate = 14.4%).

transition categories cannot be determined by the syntactic features alone, and then, various kinds of prosodic features contribute to the decision process.

Concerning the contribution of syntax to the decision of HOLD, our model is similar to the filter model in that some of the HOLD cases can be discriminated by syntax alone. However, our model differs from the filter model in regard to CHANGE. In the filter model, the CHANGE cases are always discriminated by the way of prosody, whereas in our model, some can be discriminated by syntax alone.

However, the order of the syntax node at the top and the prosody under it, does not suggest the primacy of syntax over prosody neither in the filter model nor in our model. This order is merely one possibility. We can also derive a schema in which prosody is located at the top. One question, therefore, arises, namely, whether some cases of CHANGE and HOLD can be discriminated by some prosodic features, or some combination of them, no matter which syntactic features co-occur. This is quite possible, in particular when considering combinations of the features, since, as we showed in the previous section, prosody as a whole contributes toward turn-taking at least as strongly as syntax. We do not have enough evidence to answer the question of exactly which combinations of prosodic features have crucial contributions, and we will leave this for a future study.

4.5.2 Backchannels

4.5.2.1 Contribution of Features to Discriminating Backchannels Categories

In the same way as turn-taking, we measure and compare the degree of contribution of each feature to discriminating backchannel categories making use of two sorts of multivariate statistical analysis techniques.

Table 4.11 summarizes the results of Quantification II. A large positive value of the normalized category score indicates a strong contribution of that feature to BC, whereas a large negative value indicates a strong contribution to NO-BC.

Concerning NO-BC, some syntactic features make extremely strong contributions, while none of the prosodic features do. By contrast, some part of speech

Table 4.11: Results of Quantification II: normalized category scores assigned to prosodic and syntactic features with regard to the discrimination of backchannel categories (correlation rate = .24, discrimination rate = 76.8%).

Feature		Normalized category score
Part of speech	= v-adv	1.92
Part of speech	= p-conj	1.37
F0 pattern	= rise	1.17
F0 pattern	= rise-fall	1.04
Part of speech	= v-cond	0.56
Part of speech	= noun	0.39
F0 pattern	= flat-fall	0.30
Peak F0	= low	0.24
Part of speech	= p-case	0.22
Part of speech	= p-intj	0.20
Duration	= long	0.17
Energy pattern	= late-decr	0.14
Part of speech	= v-euph	0.05
Peak energy	= high	0.04
Duration	= normal	−0.04
Energy pattern	= decr	−0.05
Peak energy	= low	−0.08
F0 pattern	= fall	−0.22
Part of speech	= v-imp	−0.24
Energy pattern	= non-decr	−0.25
F0 pattern	= flat	−0.34
Peak F0	= high	−0.35
Duration	= short	−0.45
Part of speech	= v-conc	−0.57
Part of speech	= p-final	−0.59
Part of speech	= adverb	−0.76
Part of speech	= filler	−0.80
Part of speech	= intj	−0.83
Part of speech	= conj	−0.91
Part of speech	= adnoun	−0.93
Part of speech	= v-attr	−1.12

Table 4.12: Error rates of X-removed decision trees predicting backchannel categories for each feature X. The increase in the error rate compared to the baseline is shown in parentheses.

	Without continuous variables	With continuous variables
Baseline	12.4%	9.8%
Part of speech	17.1% (4.7%)	14.2% (4.4%)
Duration	13.6% (1.2%)	11.6% (1.8%)
F0 pattern	14.7% (2.3%)	11.0% (1.2%)
Peak F0	13.8% (1.4%)	11.2% (1.4%)
Energy pattern	14.1% (1.7%)	12.1% (2.3%)
Peak energy	12.8% (0.4%)	10.4% (0.6%)
Whole prosody	18.2% (5.8%)	18.2% (8.4%)

values and F0 patterns have extremely strong contributions to discriminating BC. This result differs from that for turn-taking. We didn't find any decisive prosodic features for turn-taking but, for backchannels, there seem to be some decisive prosodic features which have as strong a contribution as the syntactic features to the presence of backchannels.

Next, we compared the predictive powers of sets of features, each set constructed by removing one particular feature from the whole set of features. The left column in Table 4.12 shows the results. It is found that, in parallel to the results for turn-taking, the increase in the error rate for the syntactic feature (4.7%) is larger than that for any of the individual prosodic features (1.2%, 2.3%, 1.4%, 1.7%, and 0.4%). This suggests that syntax has a stronger contribution to discriminating backchannel categories than any individual prosodic feature. The same can be said of the case where duration, peak F0, and peak energy are represented as continuous values (the right column in Table 4.12).

In order to compare the contribution of syntax and that of prosody as a whole, we obtained the error rate of the decision tree constructed from the syntactic feature only. The bottom-most row in Table 4.12 shows the results. The increase

in the error rate for the whole prosody is 5.8% (without continuous variables) and 8.4% (with continuous variables), which are as large as, or even larger than, corresponding error rates for syntax (without continuous variables: 4.7%, with continuous variables: 4.4%).

From these results, we can conclude that, in backchannels, (1) some instances of syntactic features, as well as some of prosodic features, have extremely strong contributions, and (2) in general, syntax has a stronger contribution than any individual prosodic feature, although prosody as a whole has slightly a stronger contribution than syntax.

4.5.2.2 Schema for Discriminating Backchannel Categories

Based on the results in the previous section, we investigate the way in which prosody and syntax are related to backchannels.

There are very few studies that attempt to build a model, such as the filter model, to capture the relationship between prosody and syntax with respect to backchannels. One extreme position in this respect is taken by Ward (1996), who claims that a certain kind of prosody, namely, changes in the F0 value during the speaker's speech, is enough to control the timing of backchannels produced by computers. On the basis of our results, we derive a schema for deciding the presence and the absence of backchannels, and examine the validity of such an extreme position.

Let us recall the results of Quantification II (Table 4.11). In regard to NO-BC, none of the prosodic features is related to it, while some syntactic features strongly contribute to it. This suggests that these syntactic features might be able to discriminate NO-BC, no matter which prosodic features co-occur. On the other hand, there are both prosodic and syntactic features which are strongly related to BC, meaning that both prosody and syntax come into play in the decision of BC.

These results for the decision process of whether or not backchannels occur can be summarized as follows (see Figure 4.6): (i) if a syntactic feature strongly connected with NO-BC appears, then backchannels tend not to occur, (ii) otherwise, (a) if a prosodic feature strongly connected with NO-BC appears, then

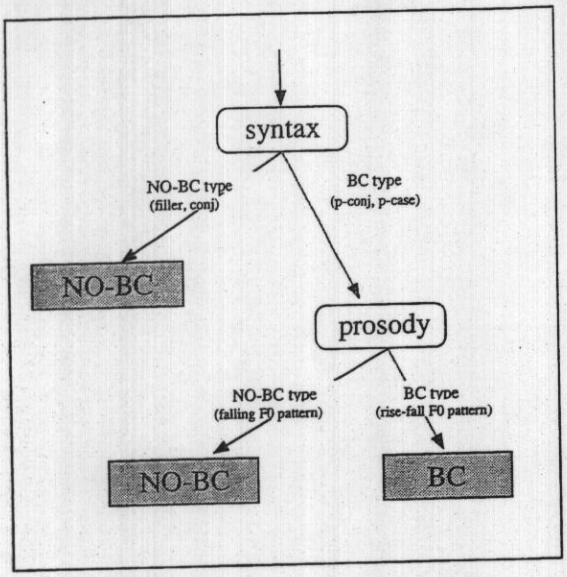


Figure 4.6: Schema for the decision process of backchannels.

backchannels tend not to occur, and (b) if a prosodic feature strongly connected with BC appears, then backchannels tend to occur. This schema is supported by the pattern of the decision tree constructed from all the features being analyzed (see Figure 4.7).

Figure 4.6 shows that, unlike our model of turn-taking, our model of backchannels is similar to the original filter model in that the absence of backchannels can sometimes be discriminated by syntax only, whereas the presence of backchannels is always discriminated by the way of prosody. The significance of prosody in this model seems to support the prosody-based approach to backchannel production like that of Ward's. However, as we mentioned before, this significance is brought by prosody as a whole, and not by any individual prosodic feature. In addition, our model suggests the importance of syntax in deciding the absence of backchannels; syntax serves as a cue for inhibiting the production of backchannels. If the relative roles of prosody and syntax with respect to backchannels are ignored, this may give rise to serious problems in purely prosody-based approaches, at least, in the language we have been investigating.

Figure 4.7: A decision tree predicting backchannel categories from syntactic and prosodic features (error rate = 12.4%).

4.6 Summary

In this chapter, we paid our attention to the local structural elements, turn-taking and backchannels. We analyzed, on the basis of corpus data of Japanese spontaneous dialogues, their relation to prosodic and syntactic features of the speaker's speech. We focused on such features as part of speech, duration, patterns of F0 contours and energy trajectories, and peak F0 and energy values at the final part of speech segments.

In the first half of the chapter, we tried to elucidate the informational potentials of prosodic and syntactic features to turn-taking and backchannels. We first showed that the features being analyzed were all related to turn-taking and backchannels, and that the way they correlated was fairly consistent with previous studies. We, then, found strong predictive powers of these features for discriminating turn-taking categories and for discriminating backchannel categories, suggesting that prosodic and syntactic features can potentially function as contextualization cues to turn-taking and backchannels. In the second half, we investigated the interrelationship between prosody and syntax in the decision process of turn-taking and backchannels. We found that, in both turn-taking and backchannels, (1) some instances of syntactic features make extremely strong contributions, and (2) in general, syntax has a stronger contribution than any individual prosodic feature, although the whole prosody contributes as strongly as, or even more strongly than, syntax. Based on these results, we illustrated the ways in which prosody and syntax are related to turn-taking and backchannels, comparing them with the "filter model" which mention roles of prosody and syntax in the decision process of turn-taking.

Chapter 5

A Non-signaling Approach to Turn-Taking

In this chapter, following the descriptive studies presented in the previous two chapters, we develop a model of a conversant's action which tries to understand the mechanism underlying turn-taking phenomena, without supposing direct use of a signaling system. We show that our non-signaling model can account for smooth transitions of turn as well as the regularities observed in wider range of conversational interactions including simultaneous starts of talks and unusual lapses between talks.

5.1 Introduction

In the previous chapter, we have explored the relation of turn-taking to several syntactic and prosodic features of the speaker's speech. These, however, are just a descriptive account of turn-taking phenomena, and they do not make clear the actual processes taken by the conversants. In this chapter, we develop a model of a conversant's action which tries to understand the mechanism underlying turn-taking phenomena.

The following two models can be hypothesized:

Signal-based model: The speaker *encodes* a meta-level message concerning

conversational organization into the current context following a certain system of codes, and the hearer *decodes* the message from the context following the same code system. Duncan's model (Duncan & Fiske, 1977), described in Section 1.2.2, is in line with this approach.

Non-signaling model: The speaker and the hearer act depending upon the context which reflects some characteristics of the on-going speech but is not necessarily seen as embodying meta-level messages, and conversational organization emerges from these actions of the speaker and the hearer.

We compare these models from the viewpoint of the type of actions performed by the conversants.

In one view, turn-taking can be regarded as a collective act, an act performed by two or more people acting collaboratively such as shaking hands, playing a duet, and paddling a canoe together (Searle, 1990; Clark & Schaefer, 1989). Such a collective act is composed of each agent's individual act which is performed as part of the collective act. In the case of turn-taking, the speaker performs an individual act of continuing or stopping his speech, while the hearer performs an individual act of starting her speech or keeping silence. In the signal-based model, it is supposed that the speaker encodes and the hearer decodes a message directly concerning turn-taking using the same system of signaling, meaning that the speaker and the hearer co-ordinate their individual acts in order to organize turn-taking as a collective act.

In the non-signaling model, on the other hand, we do not have to suppose that contexts utilized by the conversants are directly related to turn-taking. In fact, the findings in Chapter 4 suggest that such contexts might be related to the finality of a sentence. If the speaker's and the hearer's acts of starting, continuing, and stopping their speech are merely related to the finality of a sentence, these individual acts would not need collaborative coordination. In this case, turn-taking can be regarded as emerging from the speaker's and the hearer's individual acts depending upon the same context.

Let us consider an example of people gathering at a bargain sale in crowds—when people are walking down a street, a sudden loud call of "walk up! walk up!" from the store will make the store crowded. In this example, the context is

the call of "walk up! walk up!" and the individual act is each person's going to the store. They do not perform these individual acts in collaboration with each other. Rather, they only go to the store individually following the call. Thus, the act of people gathering at the store in crowds is not a collective act in the above sense, but it simply emerges from each person's individual act of going to the store.

In the above example, the context is an external event, namely, the call from the store, whereas in the case of turn-taking, the context would probably be the speaker's speech. This yields a kind of circular situation—the conversants act depending on their actions themselves. This circularity, however, is different from a direct collaboration by the signaling system. Suppose that, in the bargain-sale example, some people are rushing to the store invited by the call but others are doing so simply because they are watching other people's rushes. In this case, the context for the latter group of people is other person's act of rushing to the store, but this group of people are not in direct collaboration with the rushing people. Thus, acting dependently on the context is different from acting in collaboration, even if the context involves a participant's action.

In this chapter, we explore the possibility of the non-signaling model as a model of a conversant's action. We attempt to construct a non-signaling model of turn-taking. We start with making the following three premises:

I. Each conversant has his or her own motive for speaking.

II. Each conversant acts depending upon the context characterized by the speaker's speech, which is not necessarily directly related to turn-taking.

III. Each conversant acts independently of his or her partner's intention to bring about a particular turn-taking state.

The conversants are assumed to inherently have motives for speaking; the motives are in either of the two forms depending on the roles of the conversants in a conversation at given moment, namely, the speaker and the hearer.

- The speaker tries to keep his chance to speak.

- The hearer watches for her chance to speak.

We assume that the speaker and the hearer, each with a different type of motive for speaking, act *dependently* on the conversational context characterized by the speaker's speech. For example, if the speaker's speech is approaching the end of a sentence, the speaker might stop speaking and the hearer might start speaking. This, however, does not mean that the speaker and the hearer together coordinate turn-taking. Rather, our premise III says that they do not. In this model, smooth transitions of speaking turn can be achieved as a result of the actions of the speaker and the hearer, which are well oriented to the context, but not necessarily attuned to their partners' intentions to bring about the next turn-taking state. In the above example, a smooth change of turn would be achieved as a result of the speaker's suspension and the hearer's start of speech, which depend on the context of the speaker's speech being approaching the sentence end.

In the following sections, we examine the validity of the latter two of our three premises, II and III, based upon an analysis of a spontaneous dialogue corpus in Japanese. We start by verifying our premise II. We investigate the characteristics of the contexts in which speaking turns are changed or held. We analyze syntactic and prosodic features of the speaker's speech around turn transitions, the same method we have developed in the previous chapter. We first focus on the relation of CHANGE and HOLD to individual features and then on the correlation between the number of CHANGE- (HOLD-)related features and the probability of CHANGE (HOLD) taking place. This establishes the contexts of turn transitions characterized by the speaker's speech. Next, in order to examine the independency of the speech actions of the speaker and the hearer, our premise III, we consider whether the CHANGE- (HOLD-)related features of the speaker's speech are associated directly with turn transitions, or with each conversant's speech/non-speech actions. If the conversants acts independently of their partners' intentions about turn-taking states, the CHANGE- (HOLD-)related features should be connected with speech/non-speech actions of individual conversants, rather than turn-taking states. Such investigation is made possible by taking into account non-smooth transitions such as simultaneous starts of talks and unusual lapses between talks. Finally, we discuss how smooth transitions of speaking turn

can be achieved by these premises, without supposing direct use of a signaling system, with additional predictions about the regularities observed in non-smooth transitions.

5.2 Conversants' Dependence on the Context

Our concern in this section is to investigate the characteristics of the context in which speaking turns are changed or held, based on analyses of syntactic and prosodic features of the speaker's speech around turn transitions. We focus first on the relation of CHANGE and HOLD to individual features and then on that to the combination of the features.

5.2.1 The Relationship between Turn-Taking and Individual Features

5.2.1.1 Methods

Materials We selected eight dialogues, by sixteen different speakers, from among the Japanese Map Task Corpus, which were the same data as in Chapter 4.

Unit of Analysis As the unit of analysis, *inter-pausal unit* (IPU) were used (see Section 2.2.1).

Turn Transition Types We classify pairs of consecutive IPUs into the following four turn transition types according to whether or not the speakers of the IPUs are the same and to the temporal relation of the IPUs (See Figure 5.1):

CHANGE: the speaker of the first IPU is not the same as that of the second IPU, and the two IPUs are not overlapping with each other.

HOLD: the speaker of the first IPU is the same as that of the second IPU.[1]

[1]Note that this turn transition type was called NO-BC in Chapter 4. Here, we use a simpler term, HOLD, because we do not take backchannels into account in the following analyses (see below).

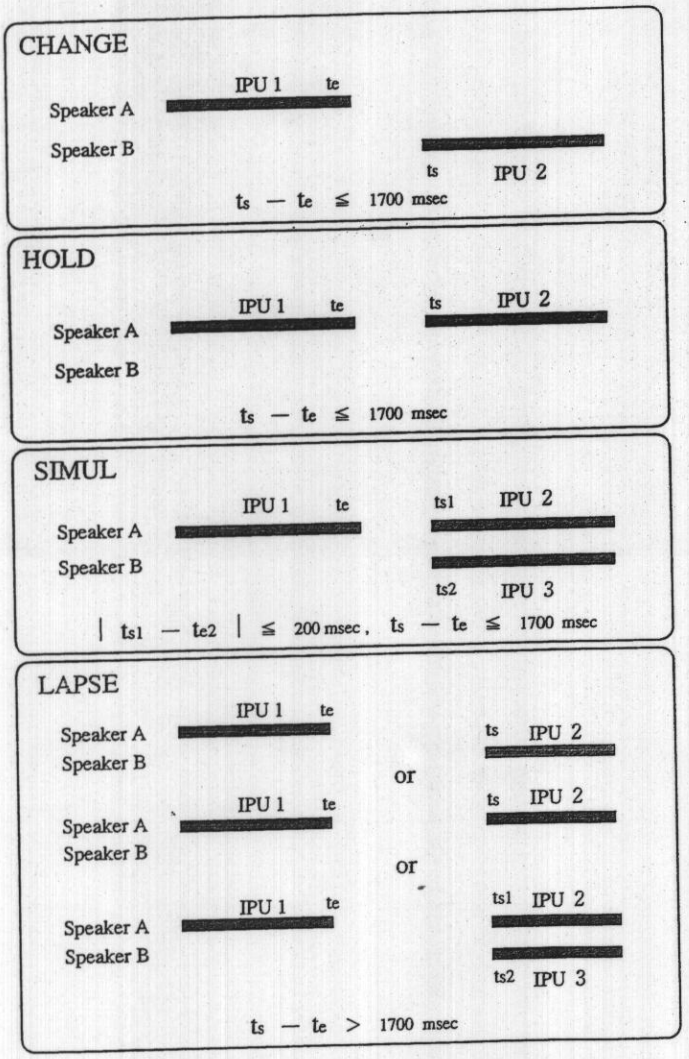


Figure 5.1: Four turn transition types.

SIMUL: the first IPU is followed by two IPUs which start almost simultaneously (the difference of the starting times is not more than 200 msec).

LAPSE: the first IPU is followed by one or two IPUs with an interval of more than 1700 msec.[2]

We left out cases where at least one IPU of the pair is a backchannel, because backchannels might affect differently on turn-taking from substantial speeches (Yngve, 1970; Schegloff, 1982; Maynard, 1989). In fact, according to the results shown in Chapter 4, almost all of the features coincident with BC are inconsistent with the features coincident with CHANGE, meaning that speaker changes accompanied by backchannels are different from speaker changes connected with substantial speeches. Two labelers identified backchannels based on their forms and functions, and cross-checked each other's results; among interjectory expressions such as "hai," "ee," and "un," those which do not constitute conversational moves were regarded as backchannels (see Section 4.2.3).

In this section, we examine only CHANGE and HOLD types, and will examine SIMUL and LAPSE types in the next section.

Syntactic and Prosodic Features We labeled each IPU with six features derived from the properties of the final portion of the IPU, which are the same ones as in Chapter 4. The six features are:

1. the part of speech of the final morpheme,
2. the duration of the final phoneme,
3. the pattern of F0 contour in the final mora region,
4. the relative height of the peak F0 of the final mora,
5. the pattern of energy trajectory in the final phoneme region, and
6. the relative height of the peak energy of the final phoneme.

[2]LAPSE type includes those cases whose configurations are the same as those of CHANGE, HOLD, and SIMUL types. We do not distinguish these three sub-types.

We assigned values of these features in the same way as in Chapter 4. As the results of the labeling, we obtained 366 IPUs for CHANGE type, and 500 for HOLD type, excluding cases whose features could not be reliably determined.

5.2.1.2 Results

Firstly, we calculated the frequencies of the syntactic and prosodic features relative to CHANGE and HOLD. Secondly, for each feature, we tested the significance of the difference of the frequencies between the two turn transition types using a binomial test, with the ratio in population being 366:500, that is the ratio of the overall frequencies of the two types. Table 5.1 shows the results. From these results, we can point out the following:

Syntactic Features:

- The features coincident with CHANGE, i.e., verbs in imperative and conclusive forms, sentence-final particles and interjections, are elements usually located at sentence final position in written Japanese, and vice versa.
- The features coincident with HOLD, i.e., conjunctions, adverbs, adnouns, case particles and filled pauses (fillers), are elements typically marking mid-sentential positions.
- The features coincident with neither CHANGE nor HOLD, i.e., conjunctive particles, nouns, and verbs in adverbial and conditional forms, are not typical of sentence final elements in written Japanese, but they can appear at sentence, or utterance, final positions in spoken Japanese when accompanied by falling or rising intonation contours.

Prosodic Features:

- The features coincident with CHANGE, such as falling and rising F0 patterns and decreasing energy patterns, are elements normally marking sentence final positions.

Table 5.1: The frequencies of features relative to CHANGE and HOLD and the results of binomial tests.

Feature	CHANGE	HOLD	Binomial test	Transition type
Duration				
short	87	**42**	CHANGE > HOLD**	TERM
normal	258	293	n.s.	NEUTRAL
long	**21**	165	CHANGE < HOLD**	KEEP
F0 pattern				
fall	**177**	104	CHANGE > HOLD**	TERM
rise	**66**	9	CHANGE > HOLD**	TERM
flat	92	**242**	CHANGE < HOLD**	KEEP
flat-fall	23	**111**	CHANGE < HOLD**	KEEP
rise-fall	8	**34**	CHANGE < HOLD**	KEEP
Peak F0				
low	**266**	286	CHANGE > HOLD**	TERM
high	100	**214**	CHANGE < HOLD**	KEEP
Energy pattern				
decr	**288**	263	CHANGE > HOLD**	TERM
late-decr	61	**175**	CHANGE < HOLD**	KEEP
non-decr	17	**62**	CHANGE < HOLD**	KEEP
Peak energy				
low	**229**	177	CHANGE > HOLD**	TERM
high	137	**323**	CHANGE < HOLD**	KEEP
Part of speech				
intj	**103**	58	CHANGE > HOLD**	TERM
v-conc	**56**	5	CHANGE > HOLD**	TERM
v-imp	**14**	3	CHANGE > HOLD**	TERM
p-final	**69**	5	CHANGE > HOLD**	TERM
noun	45	56	n.s.	NEUTRAL
v-adv	9	5	n.s.	NEUTRAL
v-euph	0	1	n.s.	NEUTRAL
v-attr	0	3	n.s.	NEUTRAL
v-cond	2	11	n.s.	NEUTRAL
p-conj	26	29	n.s.	NEUTRAL
p-intj	3	12	n.s.	NEUTRAL
adverb	3	**43**	CHANGE < HOLD**	KEEP
adnoun	0	**11**	CHANGE < HOLD**	KEEP
conj	1	**44**	CHANGE < HOLD**	KEEP
filler	2	**49**	CHANGE < HOLD**	KEEP
p-case	33	**165**	CHANGE < HOLD**	KEEP
	366	500		

(**: $p < .01$, n.s.: not significant at 1% level)

- The features coincident with HOLD, such as rise-fall and flat-fall F0 patterns and long durations, are elements marking mid-sentential positions in Japanese (Kôri, 1997; Inoue, 1997).

In summary, regardless of whether syntactic or prosodic, the features which occur when turns are changed correspond to elements marking sentence final positions, while the features which occur when turns are held are correspondent with elements marking mid-sentential positions.

5.2.2 The Relationship between Turn-Taking and the Combination of the Features

5.2.2.1 Consistency of the Combination of the Features

In the previous section, we analyzed the relation of CHANGE and HOLD to individual features. This, however, is not sufficient for characterizing the context of turn-taking, since, in speech, various features, including the six features being analyzed, appear at the same time. Therefore, we have to analyze the relation of turn transition types not only to individual features but also to the combination of the features.

It is not always the case that all the features co-occurring are consistently related to the same turn transition type. Some of them might be related to CHANGE and some to HOLD. Such inconsistency of the combination of the features seems to have relation to the probability of the occurrence of CHANGE or HOLD. For example, if we compare two cases, one in which the six features are all coincident with CHANGE and the other in which only three of the six are coincident with CHANGE, it would be expected that CHANGE is more likely to occur in the former case than in the latter, since the former contains more features coincident with CHANGE. Duncan and Fiske (1977) and Beattie (1983), in fact, pointed out that the more features coincident with CHANGE (HOLD) appear, the more frequently CHANGE (HOLD) occurs. In this section, we focus on the consistency of the combination of the features, and show that it correlates to the probability of the occurrence of CHANGE or HOLD.

5.2.2.2 Methods

Features types We classify the individual features into the following three types according to the results of the binomial test shown in Table 5.1:

TERM: features strongly associated with CHANGE (at 1% level significance)

KEEP: features strongly associated with HOLD (at 1% level significance)

NEUTRAL: features associated with neither CHANGE nor HOLD.

The ratio of TERM/KEEP features The consistency of the combination of the features is represented by the ratio of TERM (KEEP) features against all features. The ratio is given by the following formula:

$$\text{ratio of TERM (KEEP)} = \frac{\#\text{ of TERM (KEEP) features}}{\#\text{ of TERM features} + \#\text{ of KEEP features}}$$

For example, in the case where four of the six features are of TERM type and the remaining two of KEEP type, the ratio of TERM type features is $4/(4+2) = 0.67$. We suppose that NEUTRAL features do not have an influence on turn transitions. Thus, for instance, in the case where TERM, NEUTRAL, and KEEP features each appear twice, the ratio is $2/(2+2) = 0.5$.

5.2.2.3 Results

We first calculated the ratio of TERM features for each IPU, and then, for each group of IPUs whose ratios are the same, we calculated the probability of the occurrence of CHANGE in that group (= CHANGE/(CHANGE+HOLD)). We found a strong correlation between the ratio of TERM features and the probability of the occurrence of CHANGE (Figure 5.2, Spearman rank correlation: $r_s = .92$, $n = 13$, $p < .01$, two-tailed). This result shows that the more features of TERM type co-occur, the more frequently speaking turns are changed, and that, conversely, the more features of KEEP type co-occur, the more frequently speaking turns are held.

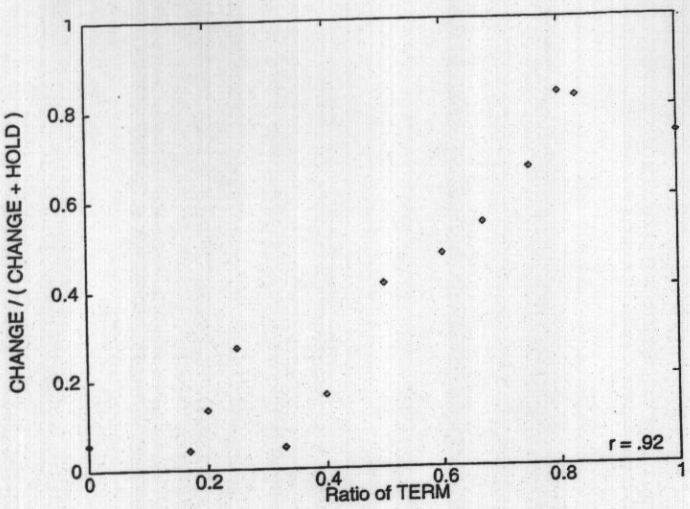


Figure 5.2: Correlation between the ratio of TERM features and the probability of the occurrence of CHANGE compared to HOLD.

5.2.3 Summary

In this section, we analyzed the relation of CHANGE and HOLD to prosodic and syntactic features in order to characterize the context of turn-taking. In Section 5.2.1, we found that the features which occur when turns are changed correspond to elements marking sentence final positions, while the features which occur when turns are held correspond to elements marking mid-sentential positions. This means that changes of turns tend to occur at the sentence boundaries, or at semantic breaks, and that they tend not to occur within a sentence, or within semantic continuity. In Section 5.2.2, we found that the more features of TERM type co-occur, the more frequently speaking turns are changed, and that the more features of KEEP type co-occur, the more frequently speaking turns are held.

These results in Sections 5.2.1 and 5.2.2 suggest that it is the context characterized by the strength of breaks of the contents of the speaker's speech that turn-taking behaviors of the conversants depend on.

5.3 Conversants' Independence of Partners' Intentions

5.3.1 Transitions of Speaking Turns or Speech/Non-Speech Actions of Conversants

In the previous section, we analyzed the relationship between the features of the speaker's speech and two smooth transition types, namely, CHANGE and HOLD. This might seem to indicate the presence of the signals directly regulating transitions of speaking turn. However, if we look at the process of turn-taking more closely, a different picture will come out.

As a matter of course, turn transitions are composed of *speech/non-speech actions* of the conversants; that is, CHANGE occurs when the previous speaker stops speaking and the previous hearer starts speaking, while HOLD occurs when the previous speaker continues speaking and the previous hearer does not start speaking (see Figure 5.3). In previous studies, only turn transitions themselves have been studied (Duncan & Fiske, 1977; Beattie, 1983; Orestrom, 1983; Ford & Thompson, 1996), and, thus, we do not have data to answer the question whether the CHANGE- (HOLD-)related features of the speaker's speech are *directly* related to turn transitions, or they are merely *indirectly* related to turn transitions, that is, such relevance of the features to turn transitions is a consequence of their relevance to speech/non-speech actions of the speaker and the hearer.

This question is very important for us to argue independency of speech/non-speech actions of the conversants, because if the conversants are acting independently of their partners' intentions to bring about a particular turn-taking state, the CHANGE- (HOLD-)related features of the speaker's speech should be connected with speech/non-speech actions of individual conversants, not with turn transitions themselves.

In the previous section, we found that there is a correlation between the ratio of TERM/KEEP features and the occurrence of CHANGE/HOLD (the diagonal arrow in Figure 5.3). If the conversants act independently of their partners' intentions, then we will find the following two correlations concerning speech/non-speech ac-

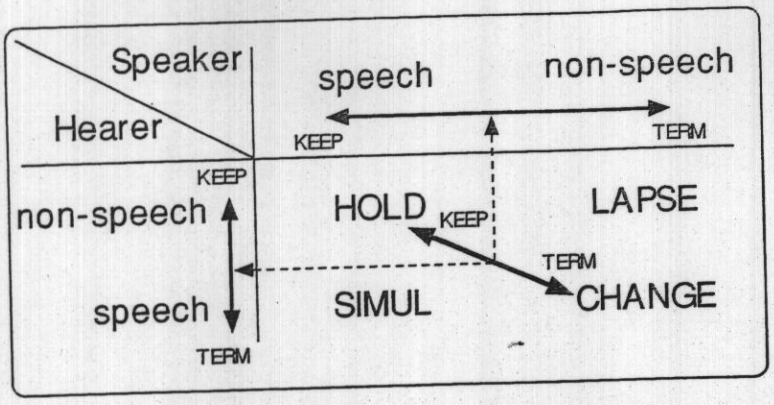


Figure 5.3: The relation between turn transitions and speech/non-speech actions of two conversants.

tions of the speaker and the hearer, which give rise to the correlation about turn transitions; one between the ratio of KEEP/TERM features and the occurrence of speech/non-speech actions of *the speaker* (the horizontal arrow in Figure 5.3) and the other between the ratio of KEEP/TERM features and the occurrence of speech/non-speech actions of *the hearer* (the vertical arrow in Figure 5.3). In this section, we show that there are indeed such correlations concerning speech/non-speech actions of the speaker and the hearer, suggesting that the conversants do not directly regulate turn-taking but act independently of their partners' intentions about turn-taking states.

5.3.2 Line of Analysis

To investigate the relation between the features of the speaker's speech and speech/non-speech actions of the speaker and the hearer, we make use not only of CHANGE and HOLD but also of SIMUL and LAPSE.

As shown in Figure 5.3, SIMUL and LAPSE are also composed of speech/non-speech actions of the two conversants, that is, SIMUL occurs when the previous speaker continues speaking and the previous hearer starts speaking, and LAPSE occurs when the previous speaker stops speaking and the previous hearer does not

start speaking. If our assumption that the CHANGE- (HOLD-)related features of the speaker's speech are connected with speech/non-speech actions of individual conversants is correct, we would find some correlations between the features and the occurrences of SIMUL and LAPSE.

Let us focus on the situation in which the hearer starts speaking (Figure 5.4 A), that is, the situation in which there are SIMUL and CHANGE. In this situation, the speech of the current speaker corresponds to SIMUL and the non-speech to CHANGE. If KEEP and TERM is related to speech/non-speech actions of the speaker, rather than turn-taking states, it is hypothesized that SIMUL occurs more often, compared to CHANGE, when the speaker's speech contains more KEEP features. That is, there is a positive correlation between the ratio of KEEP features and the probability of the occurrence of SIMUL within this situation (hypothesis A). Similarly, in the situation where the hearer does not start speaking (Figure 5.4 B), the speech of the current speaker corresponds to HOLD and the non-speech to LAPSE. If KEEP and TERM is related to speech/non-speech actions, there would be a correlation between the the ratio of KEEP features and the probability of the occurrence of HOLD within this situation (hypothesis B).

On the other hand, in the situations where the speaker continues speaking (Figure 5.5 C) and where the speaker stops speaking (Figure 5.5 D), the speech of the current hearer corresponds to SIMUL and CHANGE and the non-speech corresponds to HOLD and LAPSE, respectively. If KEEP and TERM is related to speech/non-speech actions of the hearer, it is hypothesized that the higher the ratio of TERM features is, the higher the probability of the occurrences of SIMUL and CHANGE is within these situations (hypotheses C and D).

To test the above four hypotheses, for each of the four situations (A)-(D), we analyze the relation between the consistency of the combination of the features and the two turn transition types relevant to that situation.

5.3.3 Methods

Materials The data for CHANGE and HOLD is the same ones as in Section 5.3. Since, in these eight dialogues, there are not enough amount of IPUs of SIMUL and LAPSE types for reliable analyses, for each of the sixteen speakers of the

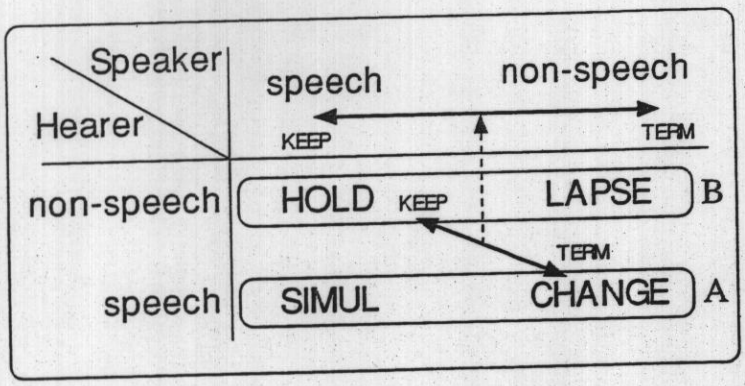


Figure 5.4: The relation between the consistency of the combination of the features and speech/non-speech actions of the previous speaker.

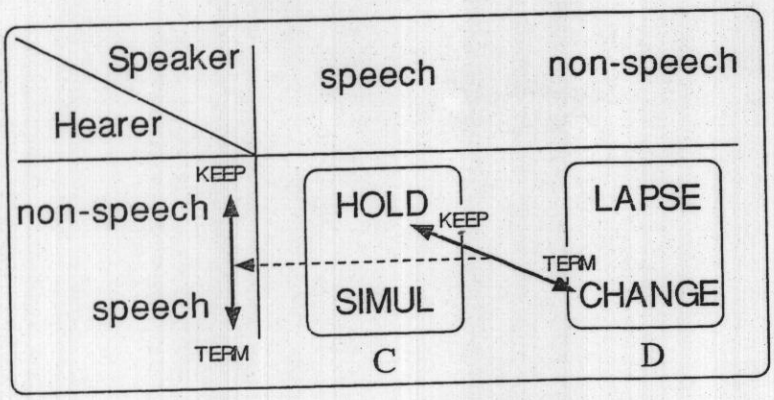


Figure 5.5: The relation between the consistency of the combination of the features and speech/non-speech actions of the previous hearer.

dialogues, we selected additional three dialogues conducted by that speaker to increase the data for SIMUL and LAPSE.[3]

Turn Transition Types SIMUL and LAPSE are defined as in Figure 5.1. SIMUL is the case where both conversants simultaneously start speaking after the previous speaker stopped his speech, while LAPSE is the case in which either or both of the conversants start speaking with a noticeable delay after the previous speaker stopped speaking. We regard a "noticeable delay" as an interval between IPUs which is much longer than usual. For each of the sixteen speakers, we calculated the mean + 1.5 standard deviation value of intervals preceded by IPUs of that speaker, for CHANGE and HOLD, and obtained 1670 msec as the average of those values for the sixteen speakers.[4] Thus, we classified into LAPSE the cases with intervals of more than 1700 msec.

Syntactic and Prosodic Features We labeled each IPU of SIMUL or LAPSE type with the same features as in Section 5.2. As the results of the labeling, we obtained 162 IPUs for SIMUL type, and 161 for LAPSE type.

The ratio of the TERM/KEEP features We calculated the ratio of TERM (KEEP) features for each IPU in the same way as in Section 5.2. We also classified each IPU into the following three groups according to the ratio of TERM/KEEP features of that IPU:

TERM group: the ratio of TERM features is above two thirds.

KEEP group: the ratio of KEEP features is above two thirds.

NEUTRAL group: the ratio of TERM/KEEP features is not above two thirds.

[3]That is, we analyzed four dialogues for each of the sixteen speakers. In this corpus, there are four dialogues conducted by the same speaker in the similar situations (see Chapter 2).

[4]The calculations were performed with logarithmic transformation to satisfy the normality of the distribution.

5.3.4 Results

For each of the three groups, TERM, KEEP, and NEUTRAL, we calculated the frequencies of IPUs in that group relative to the four turn transition types. Table 5.2 shows the results.

Table 5.2: Frequencies of IPUs relative to the four turn transition types.

Group	CHANGE	HOLD	SIMUL	LAPSE	Total
KEEP	25	209	42	29	305
NEUTRAL	110	226	67	92	495
TERM	231	65	53	40	389
Total	366	500	162	161	1189

5.3.4.1 Speech/Non-Speech Actions of the Speaker

First, to test the hypotheses (A) and (B), we focus on speech/non-speech actions of the speaker. As Figure 5.4 shows, in the situation where the previous hearer starts speaking, namely, case (A), the relation between SIMUL and CHANGE is parallel to that between speech and non-speech of the speaker; in the situation where the previous hearer does not start speaking, namely, case (B), the relation between HOLD and LAPSE is parallel to that between speech and non-speech of the speaker.

For case (A), we compared the ratios of SIMUL to CHANGE in the three combination groups (the third and the first columns in Table 5.2), and found that the ratios of SIMUL to CHANGE significantly differ among the three groups (KEEP: 62.7%, NEUTRAL: 37.9%, TERM: 18.7%; $\chi^2(2) = 55.8$, $p < .01$). Multiple comparison tests by Ryan's procedure showed significant differences in all pairwise comparisons. This result shows that the ratio of SIMUL gets higher in the order of TERM, NEUTRAL, and KEEP groups, indicating that the probability of the continuation of the speaker's speech gets higher in that order (see Figure 5.4 A). Then, in the same way as in section 5.2.2, we calculated the ratio of KEEP features (= KEEP/(KEEP+TERM)) for each IPU, and for each group of IPUs whose

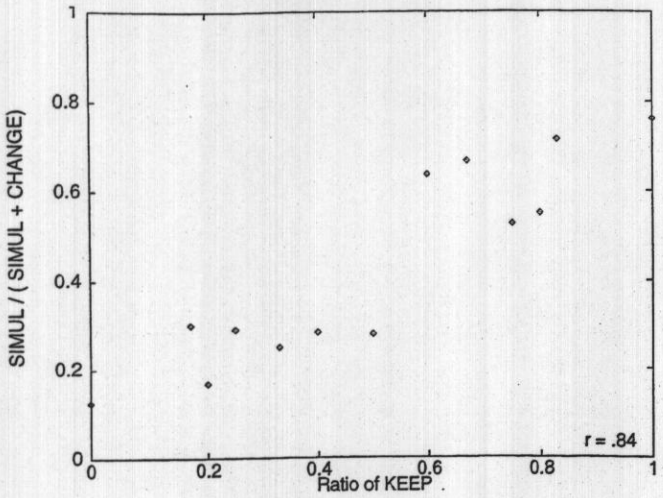


Figure 5.6: Correlation between the ratio of KEEP features and the probability of the occurrence of SIMUL in situation (A).

ratios are the same, we calculated the probability of the occurrence of SIMUL (= SIMUL/(SIMUL+CHANGE)) in that group, finding a strong correlation between the ratio of KEEP features and the probability of the occurrence of SIMUL (Figure 5.6; $r_s = .84, n = 13, p < .01$). This implies that the higher the ratio of KEEP features is, the higher the probability of the continuation of the speaker's speech is. These two results support hypothesis (A).

In the same way, for case (B), we compared the ratios of HOLD to LAPSE in the three combination groups (the second and the fourth columns in Table 5.2), and found that the ratios of HOLD to LAPSE significantly differ among the groups (KEEP: 87.8%, NEUTRAL: 71.1%, TERM: 61.9%; $\chi^2(2) = 33.5$, $p < .01$). By multiple comparison tests, significant differences were found between KEEP and TERM groups and between KEEP and NEUTRAL groups, but the difference between NEUTRAL and TERM groups was not significant. This result shows that the ratio of HOLD is higher in KEEP groups than in NEUTRAL/TERM groups, indicating that the probability of the continuation of the speaker's speech is high in KEEP group (see Figure 5.4 B). Then, we analyzed the correlation between the ratio of KEEP

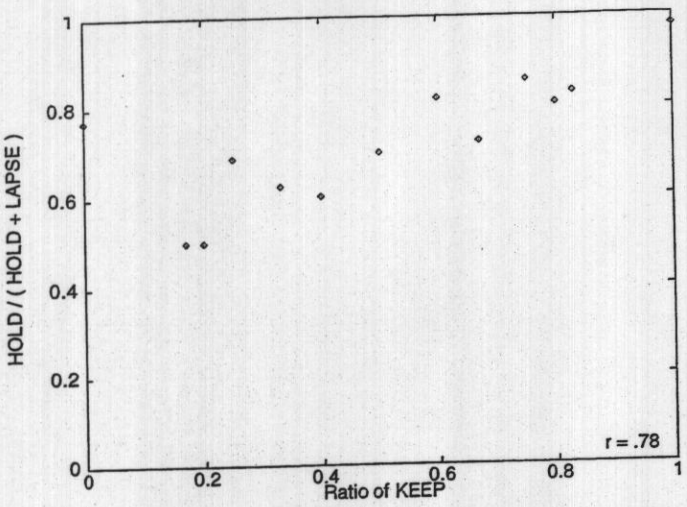


Figure 5.7: Correlation between the ratio of KEEP features and the probability of the occurrence of HOLD in situation (B).

features and the probability of the occurrence of HOLD (= HOLD/(HOLD+LAPSE)), and found a strong correlation between them (Figure 5.7; $r_s = .78$, $n = 13$, $p < .01$). This implies that the higher the ratio of KEEP features is, the higher the probability of the continuation of the speaker's speech is. These results support hypothesis (B).

In summary, these results for cases (A) and (B) show that a high ratio of KEEP features is connected to frequent occurrences of the continuation of the speaker's speech, while a low ratio to frequent occurrences of the suspension. This suggests that the consistency of the combination of the features is strongly related to speech/non-speech actions of the speaker in the way predicted in Figure 5.4.

5.3.4.2 Speech/Non-Speech Actions of the Hearers

Next, to test the hypotheses (C) and (D), we concentrate on cases (C) and (D). Here, we examine whether there is a correlation between the consistency of the combination of the features and speech/non-speech actions of the hearer (Fig-

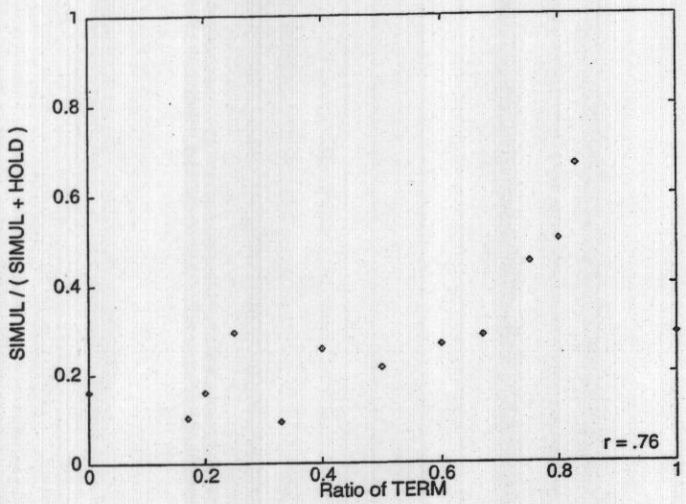


Figure 5.8: Correlation between the ratio of TERM features and the probability of the occurrence of SIMUL in situation (C).

ure 5.5).

For case (C), we compared the ratios of SIMUL to HOLD in the three combination groups (the third and the second columns in Table 5.2), and found that the ratios of SIMUL to HOLD significantly differ among the groups (KEEP: 16.7%, NEUTRAL: 22.9%, TERM: 44.9%; $\chi^2(2) = 35.2$, $p < .01$). Multiple comparison tests showed significant differences between KEEP and TERM groups and between NEUTRAL and TERM groups. Then, we analyzed the correlation between the ratio of TERM features and the probability of the occurrence of SIMUL (= SIMUL/(SIMUL+HOLD)), and found a strong correlation between them (Figure 5.8; $r_s = .76$, $n = 13$, $p < .01$). These results support hypothesis (C).

In the same way, for case (D), we compared the ratios of CHANGE to LAPSE in the three combination groups (the first and the fourth columns in Table 5.2), and found that the ratios of CHANGE to LAPSE significantly differ among the groups (KEEP: 46.3%, NEUTRAL: 54.5%, TERM: 85.2%; $\chi^2(2) = 66.9, p < .01$). Multiple comparison tests showed significant differences between KEEP and TERM groups and between NEUTRAL and TERM groups. Then, we analyzed the correlation

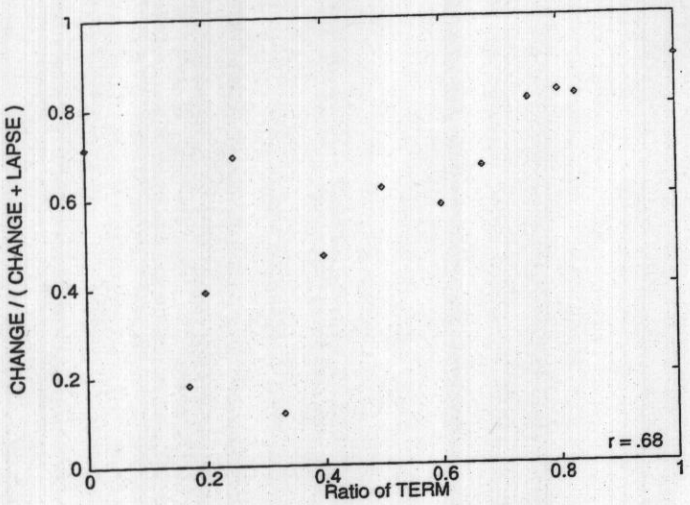


Figure 5.9: Correlation between the ratio of TERM features and the probability of the occurrence of CHANGE in situation (D).

between the ratio of TERM features and the probability of the occurrence of CHANGE (= CHANGE/(CHANGE+LAPSE)), and found a strong correlation between them (Figure 5.9; $r_s = .68$, $n = 13$, $p < .05$). These results support hypothesis (D).

In summary, these results for cases (C) and (D) show that a high ratio of TERM features is connected to frequent occurrences of the start of the hearer's speech, while a low ratio to frequent occurrences of the continuous silence. This suggests that the consistency of the combination of the features is strongly related to speech/non-speech actions of the hearer in the way predicted in Figure 5.5.

5.3.5 Summary

In this section, to test hypotheses (A)–(D), we examined the relationship between the consistency of the combination of the features and speech/non-speech actions of the conversants, finding strong correlations between them; the higher the ratio of TERM features is, the more frequently the previous speaker stops speaking

and the previous hearer starts speaking, and, conversely, the higher the ratio of KEEP features is, the more frequently the previous speaker continues speaking and the previous hearer keeps silence. These results suggest that the consistency of the combination of the features is not necessarily related to turn transitions themselves, but it is likely to be related to speech/non-speech actions of the speaker and the hearer. We can, therefore, conclude that these findings are compatible with our premise III.

5.4 Discussion

From the results of Sections 5.2 and 5.3, we can draw the following consequences:

(1) The speaker tends to stop speaking at semantic break.

(2) The hearer tends to start speaking at semantic break.

(3) The speaker tends to continue speaking within semantic continuity.

(4) The hearer tends to keep silence within semantic continuity.

What do these mean? Before answering this question, let us consider the following analogy, which is concerned with a driver wishing to move from a side road onto a main road (Beattie, 1983). If there are not enough spaces between cars on the main road, the driver would not readily to move onto the main road, or if there are even narrow spaces between cars, she might cut into the flow of traffic. Of course, if there are enough spaces, she would easily move onto the main road. Here, the driver wishing to move to the main road is the hearer in conversation, and the flow of traffic is the semantic flow of the speaker's speech. The hearer is watching for her chance to speak, looking for a "break," namely, a semantic break of the speaker's speech. When the speaker's speech comes to a semantic break, the hearer might easily start speaking, while the speaker would probably stop his speech as he has completed conveying a chunk of his messages. That is, there is a bias at semantic breaks that (1) the speaker tends to stop speaking and (2) the hearer tends to start speaking. This bias realizes smooth *changes* of speaking turns. On the other hand, when the speaker's speech is

within semantic continuity, the hearer, like the driver on the side road without enough spaces to move in, would not readily start speaking; while the speaker would continue his speech as he has not yet completed conveying the current chunk of his messages. That is, there is another bias within semantic continuity that (3) the speaker tends to continue speaking and (4) the hearer tends to keep silence, which realizes smooth *continuation* of the current speaker's turn.

These biases concerning speech/non-speech actions of the conversants account for the dominance of smooth transitions over non-smooth transitions, that is, smooth transitions such as CHANGE and HOLD occur more frequently than non-smooth transitions such as SIMUL and LAPSE. The signal-based approach has supposed that smooth transitions are achieved by conversants' obedience to signals which are *directly* related to turn-taking states. Our model, however, does not need such signals directly related to turn-taking. Smooth transitions can be arranged in great amounts of occasions as a result of the above two biases about speech/non-speech actions of the conversants, which depend merely on the context characterized by the speaker's speech, not signals specific to turn-taking.

Such different view from the signal-based approach brings about a different interpretation of the occurrence of non-smooth transitions, as well. Non-smooth transitions such as simultaneous starts of talks and unusual lapses between talks have been regarded as the phenomena caused by the neglect of signals, or rules, regulating turn-taking and, thus, as deviations from smooth transitions (Sacks et al., 1974; Levinson, 1983; Duncan & Fiske, 1977).[5] On the other hand, in Section 5.3, we showed that the occurrences of SIMUL and LAPSE, as well as CHANGE and HOLD, are dependent upon the consistency of the combination of the features of the speaker's speech, meaning that simultaneous starts and lapses do not occur contingently, but, rather, they occur under a kind of order. Such regularities on the occurrence of simultaneous starts and lapses suggest that non-smooth transitions are not deviant phenomena, and that, thus, conversants are not supposed

[5]Sacks et al. (1974) pointed out that, in conversations conducted by more than two conversants, simultaneously started turns caused by "multiple hearers" may occur, and that this kind of simultaneous starts could be explained by their model. However, our concern here is simultaneously started turns caused by the speaker and the hearer, which are regarded as deviations from smooth transitions in their model.

to regulate transitions of turns in order to avoid non-smooth transitions.[6] In this view, the difference of the frequencies between smooth transitions and non-smooth transitions is not an important matter. What is really important is that the mechanisms underlying smooth and non-smooth transitions are the same, and that this very mechanism accounts for the imbalance in their occurrences. This poses a doubt to the previous approaches, which suppose that conversants obey signals or rules "governing turn constructions, providing for the allocation of a next turn to one party, and coordinating transfer so as to minimize gaps and overlaps" (Sacks et al., 1974).

In this chapter, beginning with three premises, we have proposed a model of a conversant's action and tried to understand the mechanism of turn-taking based on this model. Not all aspects, however, could be accounted for by these premises, because the premise I, which we have not argued so far, is not always appropriate. For example, in a classroom, when a teacher puts a question to a student, the student would have to make some response to the teacher, even if he does not know the answer. Or, when you have a conversation with a stranger and an embarrassing silence occurs, you sometimes speak reluctantly, even if you have nothing to speak. These situations cannot be accounted for by conversants' inherent motives for speaking. Then, why do conversants speak in those situations? Apparently, there are some factors that force the conversants to speak when they do not have their own motives. For example, we could suppose that the hearer's speech is demanded by a social norm that the hearer have to make appropriate "responses" to "initiations" of the speaker such as questions and requests (Coulthard, 1977). Because of such a norm, the student might have to make some response to the teacher. Or if we regard a conversation as composed of incessant flows of speech, lapses would be obstacles to be removed by the conversants, which becomes another sort of pressure on the conversants to continue speaking without strong motives.

[6] Simultaneous starts and lapses may be obstacles to smooth transitions, because when they occur, the conversants usually try to remove them. For example, when two conversants simultaneously take speaking turns, either or both of them tend to stop speaking rather than both of them continue speaking. However, we should notice that this behavior of the conversants to remove obstacles does not imply that they regulate transitions of speaking turns *in advance* in order to avoid non-smooth transitions.

Hence, the model proposed in this chapter accounts for only a portion of complicated phenomena of turn-taking. However, this model introduces a very important view, that is, the smooth transitions of speaking turns can be achieved as a result of the speech/non-speech actions of the speaker and the hearer even if they act independently of their partners' intentions to bring about particular turn-taking states, provided that they depend upon the context characterized by their flows of speech. Such resultant realization of smooth transitions of speaking turns might be fundamental characteristics of turn-taking phenomena.

5.5 Summary

In this chapter, following the descriptive studies presented in the previous chapters, we developed a model of a conversant's action which tries to understand the mechanism underlying turn-taking phenomena, without supposing direct use of a signaling system. We started with making the following three premises: (I) each conversant has his or her own motive for speaking, (II) he or she acts depending upon the context characterized by the speaker's speech which is not necessarily directly related to turn-taking, and (III) he or she acts independently of his/her partner's intention to bring about a particular turn-taking state. Based upon an analysis of a task-oriented dialogue corpus in Japanese, we showed (a) that changes of turns tend to coincide with the syntactic and prosodic features marking semantic breaks, (b) that the more features coincident with turn changes appear in combination, the more frequently turn changes occur, and (c) that this tendency is not associated directly with the transitions of speaking turns, but with each conversant's speech/non-speech actions, of which the transitions are composed. These findings are compatible with the latter two of our three premises. We further discussed that smooth transitions of speaking turns can be achieved by these premises even if we do not suppose that the conversants make direct use of a signaling system, also providing additional predictions about the regularities observed in wider range of turn-taking phenomena including simultaneous starts of talks and unusual lapses between talks.

Chapter 6

Conclusion

6.1 Summary

In this dissertation, based on analyses of a Japanese spontaneous dialogue corpus, we discussed the following topics on conversational interaction:

1. The elucidation of the function of syntactic and prosodic features as contextualization cues to conversational organization, including
   - Global structural elements: discourse structures, and
   - Local structural elements: turn-taking and backchannels
2. The construction of a model of a conversant's action concerning turn-taking

In Chapter 3, we focused on the global structural elements, discourse structures. In order to elucidate the function of dynamic speech rates as contextualization cues to discourse structures, we analyzed the relation of discourse structures to dynamic speech rates. We examined corpus data of spontaneous dialogues in Japanese, and applied statistical methods to find regular correlations between local changes in the speech rate and the structures of information being expressed. We found correspondences between speech accelerations and the absence of information openings, and between speech decelerations and the presence of information openings. These results show that the dynamic speech

rates can potentially function as contextualization cues for openings and non-openings of information expressed in dialogues. Furthermore, it was found that the correlations in question hold not only for a single speaker's utterances but also for multiple speakers' sequential utterances with or without turn changes. Thus, even when there is a change of speakers during a dialogue, the subsequent speaker decelerates his speech if his speech opens up a new piece of information; otherwise, the subsequent speaker maintains the acceleration pattern established by the preceding speaker. It can be, therefore, said that, intentionally or unintentionally, the conversants collaborate with each other to maintain these regularities governing dynamic speech rates and discourse structures.

In Chapter 4, we turned our attention to the local structural elements, turn-taking and backchannels. We analyzed their relation to several syntactic and prosodic features of the speaker's speech. We focused on such features as part of speech, duration, patterns of F0 contours and energy trajectories, and peak F0 and energy values at the final part of speech segments. In order to elucidate the function of these features as contextualization cues to turn-taking and backchannels, we examined the relevance of these features to turn-taking and backchannels. We found that (1) the features being analyzed were all related to these phenomena, (2) the way they correlated was fairly consistent with previous studies, and (3) they have strong predictive powers for discriminating turn-taking categories or backchannel categories, suggesting that prosodic and syntactic features can potentially function as contextualization cues to turn-taking and backchannels. Furthermore, we investigated the interrelationship between prosody and syntax in the decision process of turn-taking and backchannels. We found that, in both turn-taking and backchannels, (1) some instances of syntactic features make extremely strong contributions, and (2) in general, syntax has a stronger contribution than any individual prosodic feature, although the whole prosody contributes as strongly as, or even more strongly than, syntax. Based on these results, we illustrated the ways in which prosody and syntax are related to turn-taking and backchannels, comparing them with the "filter model" which mention roles of prosody and syntax in the decision process of turn-taking.

In Chapter 5, following the descriptive studies presented in the previous chapters, we developed a model of a conversant's action which tries to understand the

mechanism underlying turn-taking phenomena, without supposing direct use of a signaling system. We started with making the following three premises: (I) each conversant has his or her own motive for speaking, (II) he or she acts depending upon the context characterized by the speaker's speech which is not necessarily directly related to turn-taking, and (III) he or she acts independently of his/her partner's intention to bring about a particular turn-taking state. Based upon an analysis of a task-oriented dialogue corpus in Japanese, we showed (a) that changes of turns tend to coincide with the syntactic and prosodic features marking semantic breaks, (b) that the more features coincident with turn changes appear in combination, the more frequently turn changes occur, and (c) that this tendency is not associated directly with the transitions of speaking turns, but with each conversant's speech/non-speech actions, of which the transitions are composed. These findings are compatible with the latter two of our three premises. We further discussed that smooth transitions of speaking turns can be achieved by these premises even if we do not suppose that the conversants make direct use of a signaling system, also providing additional predictions about the regularities observed in wider range of turn-taking phenomena including simultaneous starts of talks and unusual lapses between talks.

6.2 Future Direction

We believe that the present study contributes to extending our knowledge of how conversants are acting to achieve smooth interaction with access to conversational contexts. There is, however, still much to be done. We have noted several specific problems in each of Chapters 3, 4, and 5. Here, we briefly describe some of the more general steps that might to be taken.

Other Types of Contexts

In this dissertation, we focused on rather limited contexts, characterized by some sorts of syntactic and prosodic features. There remain many kinds of contexts which were not considered in this study but seem to be significant as well. For example, many researchers, using mainly English monologues, have investigated

the relation of discourse structures to linguistic features such as cue phrases and clue words (Cohen, 1984; Grosz & Sidner, 1986; Litman & Allen, 1987), and prosodic features such as pitch, intonation, energy, and pausal duration (Brown et al., 1980; Silverman, 1987; Swerts & Geluykens, 1994; Hirschberg & Nakatani, 1996; Passonneau & Litman, 1996). Then, in order to illustrate the whole picture of the influence of the conversational contexts with respect to the global structural aspects of conversation, we have to examine whether the same relationship holds or not in Japanese dialogues.

With regard to the local structural aspects such as turn-taking and backchannels, some researchers mentioned the importance of non-verbal features such as postures, gazes, and gesticulations in face-to-face conversations (Kendon, 1967; Duncan & Fiske, 1977; Goodwin, 1981), which were not taken into account in this study. These non-verbal factors have strong influence on interaction between conversants in Japanese conversations. The investigation of the non-verbal factors would, therefore, make the local structural aspects in Japanese face-to-face conversations more clear.

Other Types of Dialogues

Throughout the dissertation, we have analyzed one sort of dialogues alone, that is, task-oriented dialogues in Japanese, where two conversants participated in a conversation so as to achieve a specific goal in experimental settings. The investigation based solely on such a limited resource leaves some questions about the generality of the findings we have found as well as of the model of turn-taking we proposed. Therefore, in order to examine the generalizability of our observations and the proposed model, it is necessary to look into other styles of conversation, such as casual conversation and multi-party conversation, where more than two conversants take part. Furthermore, much insight would be gained from cross-linguistic studies.

The Relation of Turn-Taking to Global Structural Aspects

In Chapter 5, we developed a new model of turn-taking. This model, however, takes into account only local contexts, that is, the features of the speaker's speech immediately before turn transitions. Turn-taking would not be affected by the local aspects only but is also affected by global aspects of conversation such as discourse and topic structures. The elucidation of the relationship between such global aspects and turn-taking would help us to further develop our model of turn-taking.

References

Abercrombie, D. (1967). *Elements of general phonetics.* Edinburgh: Edinburgh University Press.

Anderson, A. H., Bader, M., Bard, E. G., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech, 34*, 351–366.

Aono, M., Ichikawa, A., Koiso, H., Satoh, S., Naka, M., Tutiya, S., Yagi, K., Watanabe, N., Ishizaki, M., Okada, M., Suzuki, H., Nakano, Y., & Nonaka, K. (1994). The Japanese Map Task Corpus: An interim report (in Japanese). In *Spoken language understanding and discourse processing* (Research Notes No. SIG-SLUD-9402, pp. 25–30). Japanese Society for Artificial Intelligence.

Auer, P. (1996). On the prosody and syntax of turn-continuations. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation* (pp. 57–100). Cambridge: Cambridge University Press.

Auer, P., & Luzio, A. (1992).. *The contextualization of language.* Amsterdam: Benjamins Publishing.

Ball, P. (1975). Listener's response to filled pauses in relation to floor apportionment. *British Journal of Social and Clinical Psychology, 14*, 423–424.

Beattie, G. (1983). *Talk: An analysis of speech and non-verbal behavior in conversation.* Milton Keynes: Open University Press.

Brown, G. (1977). *Listening to spoken English.* New York: Longman.

Brown, G., Currie, K., & Kenworthy, J. (1980). *Questions of intonation.* Baltimore: University Park Press.

Brubaker, R. S. (1972). Rate and pause characteristics of oral reading. *Journal of Psycholinguistic Research, 1,* 141–147.

Butterworth, B. (1974). Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research, 4,* 75–87.

Campbell, W. N. (1996). Autolabelling Japanese ToBI. In *Proceedings of the 4th International Conference on Spoken Language Processing* (pp. 2399–2402).

Campbell, W. N., & Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics, 19,* 37–47.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics, 23,* 13–31.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13,* 259–294.

Cohen, R. (1984). A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics* (pp. 251–255).

Coulthard, M. (1977). *An introduction to discourse analysis.* New York: Longman.

Den, Y., & Koiso, H. (1998). Temporal structure of conversational interaction (in Japanese). In *Proceedings of the 15th Annual Conference of the Japanese Cognitive Science Society* (pp. 160–161).

Duncan Jr., S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*(2), 283–292.

Duncan Jr., S., & Fiske, D. W. (1977). *Face-to-face interaction: Research, methods, and theory.* Hillsdale, NJ: Lawrence Erlbaum.

Edelsky, C. (1981). Who's got the floor. *Language in Society, 10*, 383–421.

Erickson, F. (1979). Talking down: Some cultural sources of miscommunication in interracial interview. In A. Wolfgang (Ed.), *Nonverbal behavior: Applications and cultural implications* (pp. 99–126). New York: Academic Press.

Erickson, F., & Shultz, J. (1981). When is a context? Some issues and methods in the analysis of social competence. In J. L. Green & C. Wallet (Eds.), *Ethnography and language in educational settings.* Norwood, NJ: Ablex Publishing Corporation.

Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and grammar* (pp. 53–12). Cambridge: Cambridge University Press.

Fujisaki, H., & Higuchi, N. (1980). Temporal organization of segmental features in Japanese disyllables (in Japanese). *The Journal of the Acoustical Society of Japan, 1*, 25–30.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers.* New York: Academic Press.

Grosz, B. J., Pollack, M. E., & Sidner, C. L. (1989). Discourse. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 437–468). Cambridge: MIT Press.

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics, 12*, 175–204.

Gumperz, J. (1989). *Contextualization cues and metapragmatics: The retrieval of cultural knowledge.* MS, University of Berkeley.

Gumperz, J. (1991). Contextualization and understanding. In A. Daranti & C. Goodwin (Eds.), *Rethinking context.* Cambridge: Cambridge University Press.

Gumperz, J., & Cook-Gumperz, J. (1982). Introduction: Language and the communication of social identity. In J. J. Gumperz (Ed.), *Language and social identity* (pp. 1–24). Cambridge: Cambridge University Press.

Haruno, M., Shirai, S., & Ôyama, Y. (1998). Using decision trees to construct a practical parser. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics.*

Hayashi, C., Diday, E., Jambu, M., & Ôsumi, N. (1988). *Recent development in clustering and data analysis.* New York: Academic Press.

Hinds, J. (1978). Conversational structure: An investigation based on Japanese interview discourse. In J. Hinds & I. Howard (Eds.), *Problems in Japanese syntax and semantics* (pp. 79–121). Tokyo: Kaitakusha.

Hirschberg, J., & Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics, 19.*

Hirschberg, J., & Nakatani, C. H. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (pp. 286–293).

Hirschberg, J., & Pierrehumbert, J. (1986). The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting of Association for Computational Linguistics* (pp. 136–144).

Horiuchi, Y., Yoshino, A., Naka, M., Tutiya, S., & Ichikawa, A. (1997). The Chiba Map Task Dialogue Corpus Project (in Japanese). In *Journal of Faculty of Engineering, Chiba University* (Vol. 2; Bulletin No. 48, pp. 33–60). Chiba University.

Imaishi, S. (1994). The use of aizuchi in natural Japanese discourse (in Japanese). In *Nihongakuho* (Bulletin No. 13, pp. 107–121). Osaka University.

Inoue, H. (1997). Sociality in intonation (in Japanese). In T. Kunihiro, H. Hirose, & M. Kohno (Eds.), *Accent, intonation, rhythm, and pause* (pp. 169–202). Tokyo: Sanseido.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22–63.

Koopmans-van Beinum, F. J., & van Donzel, M. E. (1996). Relationship between discourse structure and dynamic speech rate. In *Proceedings of the 4th International Conference on Spoken Language Processing* (pp. 1724–1727).

Kôri, S. (1996). Studies on sentence in the view of characteristics of a phonetic sound (in Japanese). *Nihongogaku, 15*(9), 60–70.

Kôri, S. (1997). Intonation in Japanese — its form and function — (in Japanese). In T. Kunihiro, H. Hirose, & M. Kohno (Eds.), *Accent, intonation, rhythm, and pause* (pp. 143–168). Tokyo: Sanseido.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics, 5*, 253–263.

Lehiste, I. (1979). Perception of sentence and paragraph boundaries. In B. Lindblom & S. Oehman (Eds.), *Frontiers of speech research* (pp. 191–201). London: Academic Press.

Levinson, S. C. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

Litman, D. J., & Allen, J. F. (1987). A plan recognition model for subdialogues in conversation. *Cognitive Science, 11*, 163–200.

Mann, W. C., & Thompson, S. A. (1986). Relational propositions in discourse. *Discourse Processes, 9*, 57–90.

Marcus, S. M. (1981). Acoustic determinants of perceptual center (p-center) location. *Perception and Psychophysics, 30*, 247–256.

Maynard, S. (1986). On back-channel behavior in Japanese and English casual conversation. *Linguistics, 24*, 1079–1108.

Maynard, S. (1989). *Japanese conversation: Self contextualization through structure and interactional management.* Ablex Publishing Corporation.

Mizutani, N. (1988). Discussion on backchannels (in Japanese). *Nihongogaku*, *7*(12), 4–11.

Nakajima, S., & Tsukada, H. (1997). Prosodic features of utterances in task-oriented dialogues. In Y. Sagisaka, W. N. Campbell, & N. Higuchi (Eds.), *Computing prosody: Computational models for processing spontaneous speech* (pp. 81–93). New York: Springer-Verlag.

Nakano, Y., Naka, M., Horiuchi, Y., Yoshino, A., Tutiya, S., Ichikawa, A., Ishizaki, M., Okada, M., Koiso, H., & Suzuki, H. (1997). Basic statistics of the Japanese Map Task Corpus (1) (in Japanese). In *Spoken language understanding and discourse processing* (Research Notes No. SIG-SLUD-9701, pp. 19–24). Japanese Society for Artificial Intelligence.

Noguchi, H. (1998). *Automatic detection of the contexts of backchannel responses based on prosodic cues.* Unpublished master's thesis, Graduate School of Information Science, Nara Institute of Science and Technology.

Orestrŏm, B. (1983). *Turn-taking in English conversation.* Lund: CWK Gleerup.

Osaka, N. (1988). The surface level exchange model in real conversation using Petri Net (in Japanese). In *Natural language understanding and models of communication* (Research Notes No. SIG-NLC88-21, pp. 57–64). The Institute of Electronics, Information and Communication Engineers.

Passonneau, R. J., & Litman, D. J. (1996). Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence, and linguistic devices. In E. H. Hovy & D. R. Scott (Eds.), *Computational and conversational discourse* (pp. 161–194). Berlin: Springer Verlag.

Passonneau, R. J., & Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, *23*, 103–139.

Pierrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In O. R. Cohen, J. Morgan, &

M. E. Pollack (Eds.), *Intentions in communication* (pp. 271–311). Cambridge, MA: MIT Press.

Pike, K. (1945). *The intonation of American English.* Ann Arbor: University of Michigan Press.

Quinlan, R. J. (1992). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann Publisher.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696–735.

Sagisaka, Y., & Tôkura, Y. (1984). Phoneme duration control for speech synthesis by rule (in Japanese). *The Transactions of the Institute of Electronics, Information and Communication Engineers, J67-A*, 629–636.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some use of "uh-huh" and other things that come between sentences. In D. Tannen (Ed.), *Analyzing discourse: Text and talk* (pp. 71–93). Georgetown University Press.

Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and grammar* (pp. 52–133). Cambridge: Cambridge University Press.

Schegloff, E. A., & Sacks, H. (1973). Opening up closing. *Semiotica, 8*, 289–327.

Searle, J. R. (1990). Collective intentions and actions. In O. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 401–415). Cambridge, MA: MIT Press.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours.* Unpublished doctoral dissertation, Cambridge University.

Stenström, A. (1994). *An introduction to spoken interaction.* New York: Longman Publishing.

Sugitô, M. (1989). Pause and intonation in discourse (in Japanese). In *The Japanese phonology* (pp. 343–364). Tokyo: Meijishoin.

Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech, 37*, 21–43.

Swerts, M., & Ostendorf, M. (1997). Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication, 22*, 25–42.

The National Language Research Institute. (1960). *A research for making sentence patterns in colloquial Japanese (1): On materials in conversation.* Tokyo: Shuuei Shuppan.

Traum, D. R. (1994). *A computational theory of grounding in natural language conversation.* Unpublished doctoral dissertation, University of Rochester.

Ward, N. (1996). Using prosodic clues to decide when to produce back-channel utterances. In *Proceedings of the 4th International Conference on Spoken Language Processing* (pp. 1728–1731).

Wightman, C. W., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing, 2*, 469–481.

Yasuda, S., & Umino, M. (1977). *Statistics in social sciences (in Japanese).* Tokyo: Maruzen.

Yngve, V. H. (1970). On getting a word in edgewise. In *the sixth regional meeting Chicago Linguistics Society* (pp. 567–577).

List of Publications

List of Major Publications

1. Koiso, Hanae, Atsushi Shimojima, and Yasuhiro Katagiri (1997). Informational potentials of dynamic speech rate in dialogue. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 394–399).
2. Koiso, Hanae, Atsushi Shimojima, and Yasuhiro Katagiri (in press). Collaborative signaling of informational structures by dynamic speech rate. *Language and Speech.*
3. Koiso, Hanae, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den (in press). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task Dialogues. *Language and Speech.*
4. Koiso, Hanae and Yasuharu Den. How does the smooth speaker transition occur? — Consideration based on the analysis of a spoken dialogue corpus — (in Japanese). submitted to *Bulletin of the Japanese Cognitive Science Society.*

List of Other Publications

1. Aono, Motoko, Akira Ichikawa, Hanae Koiso, Shinji Satoh, Makiko Naka, Syun Tutiya, Kenji Yagi, Naoya Watanabe, Masato Ishizaki, Michio Okada, Hiroyuki Suzuki, Yukiko Nakano, and Keiko Nonaka (1994). The Japanese Map Task Corpus: An interim report (in Japanese). In *SIG-SLUD-9402* (pp. 25–30).

2. Koiso, Hanae and Syun Tutiya (1995). Dialogue items in Chiba University Map Task Dialogue (in Japanese). In *SIG-SLUD-9501* (pp. 17–24).
3. Tutiya, Syun and Hanae Koiso (1995). Utterance and word units and their representations in the Chiba University Map Task Dialogue Corpus (in Japanese), In *SIG-SLUD-9501* (pp. 25–32).
4. Koiso, Hanae, Yasuo Horiuchi, Syun Tutiya, and Akira Ichikawa (1995). The acoustic properties of subutterance units and their relevance to the corresponding follow-up interjections in Japanese (in Japanese). In *SIG-SLUD-9502* (pp. 9–16).
5. Koiso, Hanae (1996). Approach to dialogues (in Japanese). *Gengo*, *25*, 20–29.
6. Koiso, Hanae, Yasuo Horiuchi, Satoshi Sasaki, Aya Yoshino, Makiko Naka, Syun Tutiya, Akira Ichikawa, Masato Ishizaki, Michio Okada, Hiroyuki Suzuki, and Yukiko Nakano (1996). The environments for the creation and analysis of Chiba Map Task Dialogue Corpus (in Japanese). In *SIG-SLUD-9503* (pp. 23–30).
7. Ishizaki, Masato and Hanae Koiso (1996). Dialogue research: Today and tomorrow (in Japanese). In *SIG-SLUD-9503* (pp. 1–8).
8. Horiuchi, Yasuo, Hanae Koiso, Syun Tutiya, and Akira Ichikawa (1996). Some utterance-final syntactic and prosodic characteristics relevant to the control of the speaker shift phenomena in spontaneous spoken dialogue (in Japanese). In *96-SLP-10* (pp. 45–50).
9. Koiso, Hanae, Yasuo Horiuchi, Syun Tutiya, and Akira Ichikawa (1996). Some utterance-final syntactic and prosodic characteristics relevant to the control of the speaker shift phenomena in spontaneous spoken dialogues (in Japanese). In *NLC-95-72* (pp. 25–30).
10. Koiso, Hanae, Yasuo Horiuchi, Syun Tutiya, and Akira Ichikawa (1996). The analysis of simultaneous utterance in Map Task Dialogue corpus (in Japanese). In *SIG-SLUD-9601* (pp. 47–54).
11. Kanda, Hirokazu, Yasuo Horiuchi, Hanae Koiso, and Akira Ichikawa (1996). Analysis of self-repaired utterances in spontaneous speech (in Japanese). In *SIG-SLUD-9601* (pp. 55–62).
12. Koiso, Hanae, Yasuo Horiuchi, Syun Tutiya, and Akira Ichikawa (1996). The prediction of the termination/continuation of utterance based on some

linguistics and prosodic elements (in Japanese). In *Proceedings of the 10th Annual Conference of the Japanese Society for Artificial Intelligence* (pp. 407–410).

13. Koiso, Hanae (1996). How to make a spoken dialogue corpus (in Japanese). In *Proceedings of the 10th Annual Conference of the Japanese Society for Artificial Intelligence* (pp. 51–54).
14. Koiso, Hanae, Yasuo Horiuchi, Satoshi Sasaki, Aya Yoshino, Makiko Naka, Syun Tutiya, Akira Ichikawa, Masato Ishizaki, Michio Okada, Hiroyuki Suzuki, and Yukiko Nakano (1996). The environments for the creation and analysis of Chiba Map Task Dialogue Corpus (in Japanese). In *Proceedings of the 13th Annual Conference of the Japanese Cognitive Science Society* (pp. 188–189).
15. Matsumoto, Yuji, and Hanae Koiso (1996). Corpora in Japanese (in Japanese). *Gengo, 25*, 114–120.
16. Matsumoto, Yuji, and Hanae Koiso (1996). Corpora in English (in Japanese). *Gengo, 25*, 121–126.
17. Koiso, Hanae, Atsushi Shimojima, Michio Okada, and Yasuhiro Katagiri (1996). Rhythmicity in conversational interaction (in Japanese). In *SIG-SLUD-9602* (pp. 19–25).
18. Koiso, Hanae (1996). How to make a spoken dialogue corpus (in Japanese). *Humane Studies and Information Processing, 12*, 21–26.
19. Indoh, Sanae, Hanae Koiso, Atsushi Shimojima, Michio Okada, and Yasuhiro Katagiri (1997). Analysis of Human Communication Mediated by Video Image (in Japanese). In *HCS96-42* (pp. 27–34).
20. Shimojima, Atsushi, Yasuhiro Katagiri, and Hanae Koiso (1997). Score-keeping in conversation-construction. In A. Benz and G. Jaeger (Eds.), *Proceedings of the Munich Workshop on Formal Semantics and Pragmatics of Dialogue MunDial'97* (pp. 172–194).
21. Nakano, Yukiko, Makiko Naka, Yasuo Horiuchi, Aya Yoshino, Syun Tutiya, Akira Ichikawa, Masato Ishizaki, Michio Okada, Hanae Koiso, and Hiroyuki Suzuki (1997). Basic Statistics of the Japanese Map Task Corpus (1) (in Japanese). In *SIG-SLUD-9701* (pp. 19–24).
22. Araki, Masahiro, Akira Ichikawa, Tatsuya Aoyagi, Masato Ishizaki, Toshihiko Itoh, Hideki Kashioka, Tomoko Kumagai, Hanae Koiso, Masafumi

Tamoto, Syun Tutiya, Shu Nakazato, Yasuo Horiuchi, Kikuo Maekawa, Yudai Murakami, Yoichi Yamashita, and Takashi Yoshimura (1998). Progress report of the discourse tagging working group (in Japanese). In *SIG-SLUD-9701* (pp. 31–36).

23. Indoh, Sanae, Noriko Suzuki, Hanae Koiso, Kazuo Ishii, Michio Okada, and Yasuhiro Katagiri (1997). Recording and analysis of the conversational corpus based on joint remembering task. In *Proceedings of the 11th Annual Conference of the Japanese Society for Artificial Intelligence* (pp. 64–65).
24. Koiso, Hanae, Atsushi Shimojima, and Yasuhiro Katagiri (1997). Information boundary and speech rate in Japanese dialogue (in Japanese). In *Proceedings of the 14th Annual Conference of the Japanese Cognitive Science Society* (pp. 242–243).
25. Swerts, Marc, Hanae Koiso, Atsushi Shimojima, and Yasuhiro Katagiri (1997). Echoing in Japanese conversations. *ATR Technical Report, TR-IT-0232.*
26. Koiso, Hanae, and Yasuharu Den (1997). An analysis of simultaneous starts of talks on the basis of the speech/non-speech configuration of conversant (in Japanese). In *SIG-J-9701* (pp. 25–30).
27. Koiso, Hanae and Yasuharu Den (1998). The relationship of syntactic and prosodic features to turn-taking in Japanese conversation (in Japanese). In *the 40th Meeting of Kinki Society of Phonetics.*
28. Ichikawa, Akira, Masahiro Araki, Masato Ishizaki, Shuichi Itabashi, Toshihiko Itoh, Hideki Kashioka, Keiji Kato, Hideki Kikuchi, Tomoko Kumagai, Akira Kurematsu, Hanae Koiso, Masafumi Tamoto, Syun Tutiya, Shu Nakazato, Yasuo Horiuchi, Kikuo Maekawa, Yoichi Yamashita, and Takashi Yoshimura (1998). The current status of standardization of discourse coding schemes (in Japanese). In *SIG-SLUD-9703* (pp. 41–48).
29. Shimojima, Atsushi, Hanae Koiso, Yasuhiro Katagiri, and Marc Swerts (1998). An analysis of echoic responses based on a spoken dialogue corpus (in Japanese). In *Proceedings of the 4th Annual Meeting of the Association for Natural Language Processing* (pp. 480–483).
30. Noguchi, Hiroaki, Hanae Koiso, Yasuko Fukuda, and Yasuharu Den (1998). Automatic detection of the contexts of backchannel responses based on

prosodic cues (in Japanese). In *Proceedings of the 4th Annual Meeting of the Association for Natural Language Processing* (pp. 484–487).

31. Ichikawa, Akira, Masahiro Araki, Masato Ishizaki, Shuichi Itabashi, Toshihiko Itoh, Hideki Kashioka, Keiji Kato, Hideki Kikuchi, Tomoko Kumagai, Akira Kurematsu, Hanae Koiso, Masafumi Tamoto, Syun Tutiya, Shu Nakazato, Yasuo Horiuchi, Kikuo Maekawa, Yoichi Yamashita, and Takashi Yoshimura (1998). Standardizing annotation schemes for Japanese discourse. In *Proceedings of the 1st International Conference on Language Resource & Evaluation* (pp. 731–735).
32. Den, Yasuharu and Hanae Koiso (1998). Temporal structure of conversational interaction (in Japanese). In *Proceedings of the 15th Annual Conference of the Japanese Cognitive Science Society* (pp. 160–161).
33. Shimojima, Atsushi, Hanae Koiso, Marc Swerts, and Yasuhiro Katagiri (to appear). An informational analysis of echoic responses in dialogue. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society.*
34. Swerts, Marc, Hanae Koiso, Atsushi Shimojima, and Yasuhiro Katagiri (to appear). Echoing in Japanese conversations. In *Proceedings of the 5th International Conference on Spoken Language Processing.*
35. Shimojima, Atsushi, Yasuhiro Katagiri, and Hanae Koiso. Themes for informational models of conversation. submitted to *Journal of Natural Language Processing.*

Abbreviations

SIG-NLC The Institute of Electronics, Information and Communication Engineers, Special Interested Group on Natural Language Understanding and Models of Communication

SIG-SLP Information Processing Society of Japan, Special Interested Group on Spoken Language Processing

SIG-SLUD Japanese Society for Artificial Intelligence, Special Interested Group on Spoken Language Understanding and Discourse Processing