DISCOURSE STRUCTURE AND ATTENTIONAL SALIENCE EFFECTS ON JAPANESE INTONATION

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of the Ohio State University

By
Jennifer J. Venditti, M.A.

* * * * *

The Ohio State University
2000

Approved by

Dissertation Committee:
Mary E. Beckman, Adviser
Michael Broe
Keith Johnson

____
Adviser
Department of Linguistics

ABSTRACT

This thesis investigates the role that intonation can play in cueing discourse structure and attentional salience in spoken Japanese. Many studies of English and other languages have investigated how the marking of intonational prominence using pitch accents is related to the attentional salience of discourse entities. In Japanese, in contrast, the presence/absence of a pitch accent is an inherent property of a word itself, and does not have such a discourse function. However, this thesis presents data suggesting that Japanese speakers can use another means to achieve the same goal: variation of pitch range can systematically mark the salience of entities in Japanese discourse.

Data were collected from a read speech database, and analyzed in terms of two well-known and widely-used theories of intonation and discourse: the Japanese ToBI intonation model, and the model of discourse structuring and global/local attentional state proposed by Grosz and colleagues. Results indicate that Japanese speakers can mark discourse structuring, global and local discourse salience, and local salience relations by varying the intonational prominence (via pitch range manipulation) of referring expressions. Specifically, speakers tend to realize discourse segment-initial phrases with high pitch range, and final phrases with lower range. In addition, Japanese speakers use intonational means to cue the global salience of discourse entities: referring expressions whose antecedent is in the current discourse segment, or a hierarchically- or linearly-recent segment have a lower range, while those which are new to the discourse or whose antecedent was in a non-recent segment tend to have a higher range. Local attentional salience and salience relations are also found to affect intonational prominence: speakers mark the local Center of attention using non-prominent intonation (all else equal). Maintenance of the Center across adjacent utterances results in non-prominent marking, while shifting the attention to a new discourse referent is marked by prominent intonation. In addition, half of the speakers adopt a strategy whereby Centering transition type interacts with the hierarchical structure of the discourse. The data suggest that Japanese speakers can manipulate pitch range to cue many of the same discourse properties which are cued by pitch accent in English.

Dedicated to the memory of
my father
James George Venditti
and his father
Thomas Venditti

ACKNOWLEDGMENTS

Many people have provided me with generous help and inspiration during my time as a graduate student. First and foremost, I would like to thank my adviser, Mary Beckman, for years upon years of guidance. From the very early days of tackling acoustic phonetics in the old Cunz Hall lab, Mary has continuously pointed me in just the right directions, showed me new challenges at every turn, and has been a trusted partner in our often extended discussions, both linguistic and otherwise. I am inspired by Mary's unending excitement about data, and by her incredible energy and desire to learn more and more (and more).

A special thanks also goes to Jan van Santen, who provided new angles and tossed many exciting challenges my way during our work together at Bell Labs. Jan helped expand my conception of intonation and acoustic modeling, speech synthesis, and of corporate life in general. I would also like to give many thanks to the members of my committee: to Keith Johnson, for always being extremely encouraging, and for giving me much needed insights and advice on all of my work in phonetics over the years; to Mike Broe, for helping me better understand the intricacies of linguistic modeling; and to Marjorie Chan, for providing many valuable comments as graduate school representative during my generals and thesis exams.

For getting me started on this path long ago, I would like to thank my mentors. Geraldine Legendre introduced me to linguistics as a science, and fueled my excitement early on. Steve Snyder went out of his way to provide individual instruction and unending support in my last year of undergraduate studies. Charles Quinn encouraged me to explore linguistics, and to go beyond the bounds of Japanese.

For many, many stimulating discussions over the years, I owe a special debt of gratitude to a number of other people. Yuki Hirose has provided countless hours of native speaker help and friendship. Julia Hirschberg has always pointed me to the right sources, and has shown continued curiosity and excitement in my research. Sun-Ah Jun has helped me develop a cross-linguistic viewpoint on intonation, and has always been a willing ear to bounce ideas off of. Kikuo Maekawa has continually provided native speaker intuitions, critiques, and encouragement. I am inspired by his dedication to opening up Japanese phonetic research to the world. I am also grateful for the extended discussions, constructive comments, and unending support from Mariapaola D'Imperio, Janet Fodor, Rebecca Herman, Stefanie Jannedy (who watched with me as the magnolias bloomed year after year), Bernd Möbius, Christine Nakatani, Marc Swerts, and Hiroko Yamashita Butler. I have learned, and continue to learn, a great deal from you all.

For extended support, challenges, and for expanding my horizons, I would like to thank all of my colleagues and mentors whom I've had the opportunity to work with over the past decade: members of the OSU Linguistics Department and Phonetics lab, the Bell Labs Language Modeling and Dialog Systems Departments, the ATR speech synthesis group, CUNY Linguistics, and OSU East Asian Studies. A special thanks goes to Kim Ainsworth-Darnell, Gayle Ayers Elam, Allison Blodgett, Julie Boland, Di Bradley, Jennifer Chu-Carroll, Serafin Coronel-Molina, Barbara Gili Fivela, Svetlana Godjevac, Jamie Green, Jen Hay, Toshio Hirai, Beth Hume, Kazumi Ikeda Banzon, Bob Kasper, Andreas Kathol, George Kiraz, Bob Levine, Joyce McDonough, Julie McGory, Kazu Maeda, Matt Makashay, Laurie Maynell, Bettina Migge, Nasiombe Mutonyi, Jim Nemcek, Amanda Miller-Ockhuizen, Ben Munson, Dave Odden, Ho-Hsien Pan, Frederick Parkinson, Jignesh Patel, Yuriy Pavlov, Shu-hui Peng, Craige Roberts, Alex Hoobie Schott, Richard Sproat, Liz Strand, Michael Tanenblatt, Moto Ueyama, Pauline Welby, Steve Winters, Shigemi Yamato, Kiyoko Yoneyama, and Suzy Bahls Yoshida. I am grateful for all of your help and support. I also owe thanks to Norio Higuchi and Nick Campbell for funding my work on Japanese ToBI during my stay at ATR, and to Joe Olive for funding work on Japanese intonation modeling for speech synthesis and for providing the facilities and resources to enable work on this thesis. The thesis research was funded in part by the OSU Department of Linguistics, and by a Graduate Student Alumni Research Award and Presidential Fellowship from the OSU Graduate School.

This thesis would not be possible without the native speaker talents and unending patience of a number of people. Yuki Hirose, Yuko Okutsu Harako, and Moto Ueyama went out of their way to provide hours upon hours of native speaker insights and help over many years. I have also learned a great deal about Japanese intonation with the help of Mari Fujimoto, Noriko Fujioka, Kaori Furuya, Takako Katsuno, Miho Mochizuki, Azusa Morii, Kayou Nagata, Seiko Nakamura, Rumiko Sode, Sonoko Takahashi, and other experiment participants.

Finally, I would like to thank my family for their continued support. My mother, Virginia Venditti, has always been there for me, and always believed it would happen someday. A very special thanks goes to my husband, Sean Ramprashad, for his constant encouragement, for giving me strength and motivation, and for reminding me what is important.

VITA

1969	Born in Syracuse, New York
1991	B.A. in Oriental Languages and Literatures, University of Colorado at Boulder
1993	M.A. in East Asian Languages and Literatures, Ohio State University
1999	Presidential Fellowship, Ohio State University
2000	Ph.D. in Linguistics, Ohio State University
2000 – present	Consultant. Lucent Technologies Bell Laboratories, Language Modeling Research Department. Murray Hill, NJ. (Project: Japanese text-to-speech synthesis and intonation modeling)

PUBLICATIONS

Venditti, Jennifer J. (forthcoming) The J_ToBI model of Japanese intonation. In S.-A. Jun (ed.) *Prosodic Typology and Transcription: A Unified Approach.* (Collection of papers from the ICPhS 1999 satellite workshop on "Intonation: Models and ToBI Labeling". San Francisco, California).

Venditti, Jennifer J. and Jan P. H. van Santen. 2000. Japanese intonation synthesis using superposition and linear alignment models. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP).* Beijing, China.

Beckman, Mary E. and Jennifer J. Venditti. 2000. Tagging prosody and discourse structure in elicited spontaneous speech. In *Proceedings of the Science and Technology Agency Priority Program Symposium on Spontaneous Speech: Corpus and Processing Technology*, pp. 87–98. Tokyo, Japan.

Venditti, Jennifer J., Kazuaki Maeda and Jan P. H. van Santen. 1998. Modeling Japanese boundary pitch movements for speech synthesis. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pp. 317–322. Jenolan Caves, Australia.

Venditti, Jennifer J. and Jan P. H. van Santen. 1998. Modeling segmental durations for Japanese text-to-speech synthesis. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pp. 31–36. Jenolan Caves, Australia.

van Santen, Jan P. H., Bernd Möbius, Jennifer J. Venditti, and Chilin Shih. 1998. Description of the Bell Labs Intonation System. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pp. 293–298. Jenolan Caves, Australia.

Venditti, Jennifer J. and Jan P. H. van Santen. 1998. Modeling vowel duration for Japanese text-to-speech synthesis. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia.

Maeda, Kazuaki and Jennifer J. Venditti. 1998. Phonetic investigation of boundary pitch movements in Japanese. In *Proceedings of the 1998 International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia.

Venditti, Jennifer J. 1997. Japanese ToBI Labelling Guidelines. In K. Ainsworth-Darnell and M. D'Imperio (eds.) *Papers from the Linguistics Laboratory*. Ohio State University Working Papers in Linguistics, vol. 50: 127–162. [First distributed in 1995 at: http://ling.ohio-state.edu/Phonetics/J_ToBI/jtobi_homepage.html].

Venditti, Jennifer J. and Marc Swerts. 1996. Intonational cues to discourse structure in Japanese. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, vol. 2: 725–728. Philadelphia, Pennsylvania.

Venditti, Jennifer J., Sun–Ah Jun and Mary E. Beckman. 1996. Prosodic cues to syntactic and other linguistic structures in Japanese, Korean and English. In J. Morgan and K. Demuth (eds.) *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, pp. 287–311. Lawrence Erlbaum Associates, Inc.

Campbell, Nick and Jennifer J. Venditti. 1995. J-ToBI: An intonation labelling system for Japanese. In *Proceedings of the Autumn meeting of the Acoustical Society of Japan (ASJ)*, vol. 1: 317–318. Utsunomiya, Japan.

Venditti, Jennifer J. and Hiroko Yamashita. 1994. Prosodic information and processing of temporarily ambiguous constructions in Japanese. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, vol. 3: 1147–1150. Yokohama, Japan.

Venditti, Jennifer J. and Hiroko Yamashita. 1994. Prosodic information and processing of complex NPs in Japanese. In H. Ura and M. Koizumi (eds.) *Proceedings of the First Conference on Formal Approaches to Japanese Linguistics.* MIT Working Papers in Linguistics, vol. 24: 375–391. Cambridge, Massachusetts.

Venditti, Jennifer J. 1994. The influence of syntax on prosodic structure in Japanese. In J.J. Venditti (ed.) *Papers from the Linguistics Laboratory.* Ohio State University Working Papers in Linguistics, vol. 44: 191–223.

Venditti, Jennifer J. (ed.) 1994. *Papers from the Linguistics Laboratory.* Ohio State University Working Papers in Linguistics, vol. 44.

FIELDS OF STUDY

Major field: Linguistics

Specialization: Phonetics

TABLE OF CONTENTS

APPENDICES

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

This thesis investigates the role that intonation can play in cueing discourse structure and attentional salience in spoken Japanese. Data were collected from a read speech database, and analyzed in terms of the Japanese ToBI intonation model [Ven95, Ven99, BP86, PB88], and the model of discourse structuring and attentional state proposed by Grosz and colleagues [GS86, GJW95]. Results indicate that Japanese speakers can mark discourse structuring, global and local discourse salience, and local salience relations by varying the intonational prominence (via pitch range manipulation) of referring expressions. These results are consistent with findings of previous studies which have examined the influence of discourse structures on pitch accent distribution in English and other languages.

1.1 The need for intonation-discourse interface research

Understanding how speakers modulate their intonation over the course of spoken discourse has been an area of much theoretical and applied linguistic research over the past decades. Many approaches to both intonation and discourse structuring have been examined, in order to investigate the relation that may hold between the two linguistic structures. However, despite these many efforts, our understanding of this mapping is still in its infancy.

In the domain of natural language generation and speech synthesis, the lack of extensive knowledge about how native speakers may vary their intonation is especially apparent. The generated speech from many state-of-the-art synthesis systems strikes listeners as 'dull' or altogether awkward. This observation has prompted the leading researchers in the field, in a recent report on the status of current speech synthesis technology, to call for more extensive research: "Prosody, which includes the phrase and accent structure of speech, is one of the least developed parts of existing speech synthesis systems" [SOH99, p. 23]. In addition, as the applications which employ speech synthesis evolve from single sentence responses to full-scale dialog or messaging systems, understanding how speakers vary intonation in connected discourse is even more critical.

Fortunately, recent advances in computational and theoretical linguistics provide the resources and tools to facilitate and improve intonation-discourse research. One resource is the availability of large databases of naturally-occurring spoken discourse.

These databases provide the wealth of data required to investigate the intonation-discourse mapping, using efficient computational techniques. Another advance is robust theoretical models of both intonational organization and discourse structuring. Such models provide a principled description of the linguistic systems underlying both intonation and discourse. Databases which are systematically tagged using these models are an invaluable resource for research on the intonation-discourse interface.

1.2 Thesis overview

This thesis employs two well-known and widely-used models, one of intonation and one of discourse, to provide a backdrop against which the relation between the two structures in a constructed Japanese read speech database can be investigated. First, Chapter 2 outlines the Japanese ToBI model of intonation description [Ven95, Ven99, BP86, PB88], including a summary of the phonological contrasts which are relevant in the Tokyo Japanese intonational system, and the phonetic realization of these contrasts in surface fundamental frequency (F0) contours. This chapter also provides a comparative discussion of accent function and intonational prominence in Japanese vs. English. Chapter 3 outlines the model of discourse structuring and attentional state proposed by Grosz and Sidner [GS86], and also the model of local discourse coherence known as Centering Theory [GJW95]. These approaches together highlight the notions of global and local discourse attentional salience, and the constraints on the interpretation of referring expressions in connected discourse. Chapter 4 then presents a review of previous studies which have investigated the intonation-discourse mapping in English, Japanese, and other languages. The chapter focuses on research that has synthesized the two approaches to intonation and discourse structuring outlined in Chapters 2 and 3, and enumerates a number of open research questions about the Japanese intonation-discourse interface which warrant empirical investigation. Chapters 5–7 describe an experimental analysis of a Japanese read speech database constructed to address these open research questions. In Chapter 5, the design and methods of data collection and analysis are described. Chapter 6 gives a detailed description of the effects of discourse structure and global attentional salience observed in the data. Then, Chapter 7 details the effects of local attentional salience, salience relations, and hierarchical discourse structure on the intonational realization of referring expressions. Finally, Chapter 8 summarizes the results and relates them to previous studies of intonation-discourse mapping in English and other languages.

CHAPTER 2
JAPANESE ACCENT AND INTONATIONAL STRUCTURE

2.1 Intonation description using Japanese ToBI

Spoken language databases contain rich intonational variation that can be linguistically analyzed, given that the corpora are systematically tagged using labels marking intonational structure. One such tagging scheme is the Japanese ToBI (henceforth J_ToBI) system, used for describing phonological events in the intonation contours of Standard (Tokyo) Japanese [Ven95, Ven99]. J_ToBI is a model which follows the *tone-sequence approach* to intonation description (see [Ladd96] for an overview of different approaches), in which tonal targets are associated to specific syllables in an utterance, and the overall fundamental frequency (F0) contour of the utterance is roughly described as a linear interpolation between these targets.[1]

J_ToBI is one of the many systems subsumed under the general class of ToBI (To̲nes and B̲reak I̲ndices) models, characterized by their use of tones and break index labels to describe the intonation tunes and prosodic structure of spoken utterances. The first ToBI system was developed in the early '90s to describe the phonological intonational contrasts found in several varieties of English, including General American English and Anglo-Australian English [BE94, SBP$^+$92, PBH94]. It was based on Pierrehumbert's tone-sequence analysis of American English [Pierre80], supplemented by input from researchers studying other varieties of English, and also from those working on the perception of prosodic breaks (e.g. [POSHF91]). Since that time, a number of other ToBI systems have been developed for languages as diverse as Cantonese, Chickasaw, French, German, Japanese, Korean, Mandarin Chinese, Serbo-Croatian, etc. [Jun]. Because of its comprehensive coverage of contrastive intonational patterns, its generalization to encompass cross-linguistic intonation description, and its widespread use in many institutions worldwide, ToBI has been widely adopted as a framework within which to build standards for tagging large spoken language databases (e.g. the Boston University radio news corpus [OPSH95], Boston Directions Corpus [HN96], TRAINS corpus [HA99], ATR-ITL database [OC95], etc.).

All ToBI systems include symbolic labels for marking distinctive tonal patterns and break indices (degrees of disjuncture between sequential units), and many also

[1] Although this thesis focuses on the tone-sequence model, the results are relevant and can be interpreted using other models of Japanese intonation as well (e.g. [FS71, FH84, VvS00], etc.).

include labels for tagging disfluencies, pragmatic effects, or other site-specific information. However, even though all ToBI systems mark tones and break indices, this is not to say that the same symbolic labels are used for each system. The tags used for a given language are determined by a thorough phonological analysis of the distinctive intonation patterns and prosodic phrase units for that language alone. In this way, ToBI is not like 'an IPA (International Phonetic Alphabet) of intonation', as it may be perceived, but rather it represents only the phonological contrasts of a given language's intonation system. The Japanese ToBI scheme has as its foundation the phonological analysis developed by Beckman and Pierrehumbert [BP86, PB88], which builds on previous analyses by Poser, Haraguchi, McCawley, Kawakami, Hattori and others (e.g. [Poser84, Hara77, McCaw68, Kawa61, Hat61, Hat60]). The analysis presented in *Japanese Tone Structure* [PB88] was further revised by Venditti based on analyses of additional data, resulting in the current J_ToBI scheme [Ven95, Ven99]. In what follows, I will give a brief overview of the system (enough to provide a base for interpreting the analysis and results presented in this thesis), and refer the reader to the J_ToBI literature for further details.

2.1.1 J_ToBI accent and phrase tones

One of the fundamental contrasts in the Japanese intonational system is the distinction between accented and unaccented lexical items. Japanese is considered a *pitch accent language*, in that it uses local pitch events to mark certain syllables in the speech stream. In this way it is similar to languages such as English, which also use pitch accents in their intonational systems. However, unlike English, where accent placement is governed by a complex interaction of syntactic structure, pragmatic function, discourse organization, and so on, the presence or absence of an accent on a particular syllable in a Japanese utterance can be predicted simply by knowing what word is being uttered. That is, it is a property of the lexical item itself.[2] Take for example the minimal pair shown in Figure 2.1.

In the figure, both panels are plotted using the same F0 scale, and the vertical lines mark the end of the second mora in both panels. Here, the verb /ueru/ in the phrase *uerumono* 'something to plant' (left panel) is lexically-specified as unaccented, while that in *ue'rumono* 'the ones who are starved' (right panel) is specified as accented on the second mora.[3] The accented phrase displays a precipitous fall in pitch starting near the end of this accented mora, while the unaccented phrase lacks such a rapid

[2] Again, the discussion and data in this thesis are limited to Tokyo Japanese accent patterns. Other dialects of Japanese may have markedly different lexical patterns (e.g. Kansai [Kori87, PB88]), or may have no accented/unaccented distinction whatsoever at the lexical level (e.g. Kumamoto and Fukui [Mae90]).

[3] In the transcriptions, accented words contain an apostrophe after the vowel with which the accent is associated; unaccented words lack such a marking.

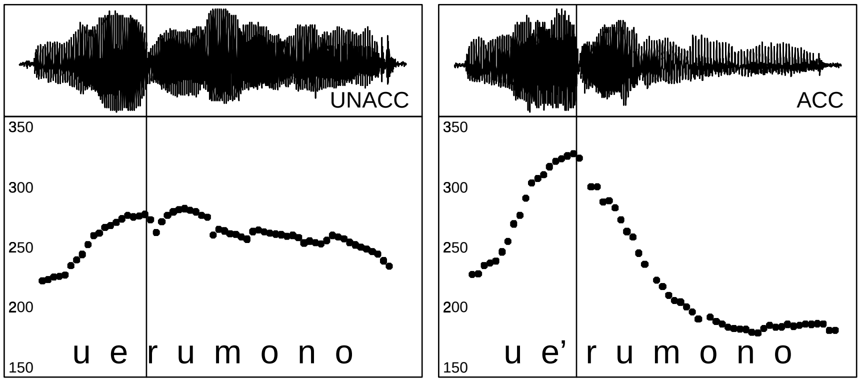

Figure 2.1: Waveforms and F0 contours of unaccented *uerumono* 'something to plant' (left) and accented *ue'rumono* 'the ones who are starved' (right) phrases, uttered by the same speaker.

fall. In the J_ToBI system, this falling F0 movement associated with the lexical accent is represented by the H*+L label: the H* starred tone indicates that the high portion is phonologically associated with the mora specified as accented in the lexicon, and the +L is a trailing low tone, which is realized shortly afterward, producing the falling pitch movement characteristic of accented words (see [PB88] for discussion of phonological association, and [VvS00] for more details of the precise temporal alignment). The unaccented phrase in the left panel of Figure 2.1, in contrast, lacks such a rapid fall. Instead, the F0 contour rises to a (shallow) peak near the end of the second mora, then gradually declines to the right phrase edge. J_ToBI marks the peak (or end of rise) in unaccented cases with a H- phrase tone, which is associated with the first or second sonorant mora of the phrase.[4]

In both the unaccented and accented phrases shown in Figure 2.1, the F0 contour rises from a low point at the left phrase edge, then declines to a low at the end of the phrase (albeit in the accented case the final low value is much lower). These low tonal targets are represented in the J_ToBI system by boundary tones %L and L%, respectively.[5] These boundary tones (and the H- phrase tone as well) are not a

[4] In cases where the first syllable is heavy (long) with a sonorant second mora, the H- associates with the first mora of the phrase. In cases where the first syllable is light (short), the H- associates with the second mora, as shown above (see [PB88], inter alia for a full discussion).

[5] There are also 'weak' variants of these boundary tones, %wL and wL%, which occur when the first syllable of the phrase (or following phrase in the case of the wL% tone) is accented or long. These variants will not be discussed further here, but the reader is referred to [PB88, Ven95] for more details.

property of the lexical item itself, but are markers of higher-level prosodic groupings of words into phrasal units, which I will describe in the following section.

2.1.2 J_ToBI prosodic phrase levels

In addition to the accented/unaccented lexical distinction, another important part of the Japanese intonation system is the grouping of words into prosodic phrases. Speakers can produce a string of words or morphemes as a single intonational unit, which is defined both tonally and by the degree of perceived disjuncture among words within and between groups. This grouping occurs at two levels in Japanese.

First, there is a lower-level grouping, such as that shown in each panel in Figure 2.1. The verb *ueru/ue'ru* is combined with the following unaccented noun *mono* 'thing/person', into a single prosodic phrase. This level of prosodic phrasing in Japanese is termed the *accentual phrase* (AP), and is typically characterized by a rise to a high around the second mora (H-), and subsequent fall to a low at the right edge of the phrase (L%). This delimitative tonal pattern is a marking of the prosodic grouping itself, separate from the F0 movement due to a pitch accent. That is, in a string of words which are combined into a single AP, the entire unit will be characterized by the delimitative rise-fall pattern (H-L%), regardless of the lexical accent specification of the component words. If all words in the phrase are unaccented, the resulting contour resembles that shown in the left panel of Figure 2.1. However, if the phrase contains an accented word, then the contour will show the signature accentual fall, in addition to the phrasal rise at the left edge.[6] The degree of perceived disjuncture between words within an accentual phrase is less than that between sequential words with an accentual phrase boundary intervening. In Tokyo Japanese it is most common for unaccented words to combine with adjacent words to form accentual phrases, though under some circumstances a sequence of accented words may combine, in which case the leftmost accent survives and subsequent accents in the phrase are deleted.

The second type of prosodic grouping in Japanese is the higher-level *intonation phrase* (IP), which consists of a string of one or more accentual phrases. As with the accentual phrase, this level of phrasing is also defined both tonally and by the degree of perceived disjuncture within/between the groups. However, the tonal markings and the degree of disjuncture are different from those of the accentual phrase. The intonation phrase is the prosodic domain within which pitch range is specified, and thus at the start of each new phrase, the speaker chooses a new range which is

[6] In cases where the accent is early in the phrase, such as on the 1st or 2nd mora (see the right panel of Figure 2.1), the target endpoint of the initial rise due to the H- will be obscured by the H* of the accent. However, if the accent is late in the phrase, the phrasal rise (H-) *and* the peak before the accentual fall (H*) will both be observed.

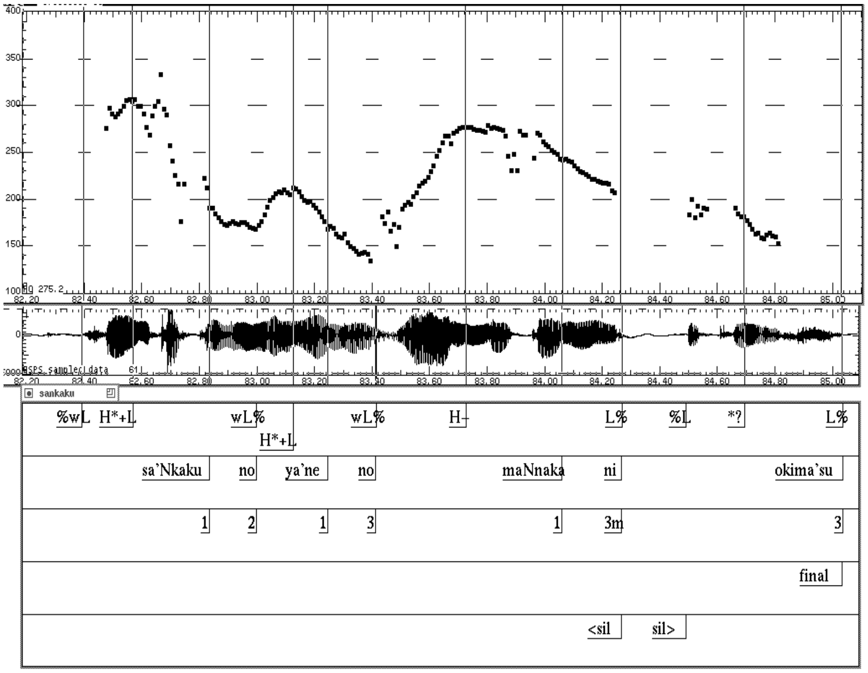

Figure 2.2: Waveform, F0 contour, and J_ToBI transcription of the utterance ≪sankaku≫: triangle-GEN roof-GEN middle-LOC put 'I will place it right in the center of the triangular roof.' From [Ven95].

independent of the former specification (see Section 2.3 below for an extended discussion). Since there also is a process of *downstep* in Japanese, by which the local pitch height of each accentual phrase is reduced when following a lexically-accented phrase, one will often observe a descending staircase-like effect of accentual phrase peaks in all-accented sequences, which is then 'reset' at an intonation phrase boundary. In addition to this behavior of the pitch range, the degree of perceived disjuncture between sequential words across intonation phrase boundaries is larger than that of words within or across accentual phrase boundaries.

Figure 2.2 contains a J_ToBI-transcribed example utterance showing words grouped into accentual phrases and higher-level intonation phrases.[7] The prosodic phrasing of this utterance was judged by a J_ToBI labeler as follows:

```
accentual phrase   {          }  {       }  {         }  {      }
intonation phrase  [                     ]  [         ]  [      ]
                   sa'nkaku  no  ya'ne  no  mannaka  ni  okima'su
```

[7] This utterance is taken from the task-based housebuilding monologue presented in [VS96].

The accented accentual phrases *sa'nkaku no* 'triangular' and *ya'ne no* 'roof-GEN' combine to form the first intonation phrase, with *ya'ne no* being downstepped due to the pitch accent on *sa'nkaku*. There is then an expansion of pitch range starting with the unaccented accentual phrase *mannaka ni* 'middle-LOC'; this and the virtual pause between *ya'ne no* and *mannaka* suggest an intonation phrase boundary.[8] In the J_ToBI transcription (see [Ven95, Ven99] for full details of all label tiers shown here), the F0 peaks due to the lexical accents in the first two accented phrases are marked in the tone tier (which is the top window of labels shown under the F0 track) by H*+L, regardless of downstepping effects. The phrasal high at the end of the rise in the unaccented phrase is marked by H-. The low AP-final boundary tones (L%/wL%) are also marked at the phrase boundaries, as shown in the figure.

In addition to the pitch range and disjuncture cues to intonation phrase boundaries, this prosodic unit is also characterized by boundary tones. The initial low tone found in post-pausal intonation phrases is marked by %L (or %wL), and serves as the tonal anchor from which the F0 rises in both panels in Figure 2.1, and in the post-pausal positions in Figure 2.2. In addition to this low marking of the left edge of the intonation phrase, there may also (optionally) be rising or rise-fall tonal movements at its right edge. These movements, marked by labels such as H%, LH% or HL%, serve to cue various pragmatic functions of the utterance, such as questioning, incredulity, explanation, insistence, etc. While investigation of the distribution and function of boundary pitch movements in discourse is crucial to our understanding of spoken language generation and processing, it does not relate directly to the role of intonation in cueing discourse salience, which is the topic of this thesis. The reader is referred to [Kawa95, VMvS98, Ven99] for extensive discussion and examples of the various boundary pitch movements occurring in Tokyo Japanese. In Figure 2.2 above, each intonation phrase ends in a low tone, without any such rising or rise-fall movement.

This section has given a brief overview of the main components of a J_ToBI prosodic analysis: the accented/unaccented lexical distinction, and the levels of prosodic phrasing, which are characterized by delimitative tonal patterns, disjuncture cues, and pitch range specification. The issues of accent function and pitch range manipulation will be discussed in more detail in the following sections.

2.2 Accent function in Japanese vs. English

As mentioned in Section 2.1.1, both Japanese and English are considered pitch accent languages, in that they both use tones to mark certain syllables in the speech stream. However, the similarity only goes so far; there are several fundamental differences in the role of pitch accents in the two languages. First, Japanese and English differ

[8] The phrasal juncture between *mannaka ni* and *okimasu* is discussed in detail in [Ven95, Ven99].

in the level (lexical vs. post-lexical) at which pitch accent comes into play in the intonational system. In Japanese, pitch accents fall on syllables specified for that property in the lexicon, as discussed above. Words are either accented or unaccented, and those which are accented have the location of the accented mora specified at the lexical level. This lexical distinction contrasts with languages such as English, in which pitch accents play a role at an entirely different level. In English, the location of metrically strong syllables in a word is determined at the lexical level, and these syllables (most often the strongest, or 'primary-stressed' syllable) serve as docking sites to which pitch accents may be associated at the post-lexical level. Thus in English, unlike in Japanese, accents are part of the metrical (rhythmic) structuring that we call 'stress'.

A second difference between the languages is the inventory of shapes and meanings of the pitch accents themselves. In Japanese there is only one type of pitch accent: a sharp fall from a high occurring near the end of the accented mora to a low in the following mora (represented by H*+L in J_ToBI). In English, the inventory of pitch accent shapes is far more diverse. There are a number of pitch accent shapes, in which the F0 can rise or fall around the accented syllable, or can maintain a local maximum/minimum on that syllable. Rising or falling accents can further be distinguished in terms of where the movement is anchored relative to the accent (i.e. whether there is a leading tone or trailing tone). Each of these accent shapes has associated with it a specific pragmatic meaning which that accent lends to the overall meaning of the intoned utterance (see [PH90], inter alia). The Japanese falling accent does not have any such meaning associated with it.

A third difference between the two languages (and probably the most relevant for this thesis), is the *function* and *distribution* of pitch accents. In English, pitch accents serve to highlight, or make 'prominent' certain words or syllables in the discourse, and the distribution of pitch accents in an English utterance reflects this function. In a given utterance, there will be a number of metrically strong syllables that can potentially be made even more prominent by the association of a pitch accent. On which of these syllables pitch accents will fall is highly dependent on the linguistic structure of the utterance. That is, a complex interaction of various factors such as the syntactic structure, semantic content, pragmatic function, discourse organization, or attentional state, etc. will determine where the pitch accents are to be placed in English (see discussion of discourse-related effects on English accent placement in Chapter 4 below). In Japanese, in contrast, pitch accent is a lexical property of a given word, and thus it (in and of itself) is not able to contribute any such prominence-lending function. This leaves little room for variability in distribution of accents in a Japanese utterance or across a discourse. However, as I will show in Chapters 4, 6 and 7, Japanese does employ other means for cueing prominence in discourse. Specifically, the data presented in this thesis suggest that Japanese uses variation in *pitch range* in much the same way as variation in accent placement is used in English, for marking the salience of discourse entities.

2.3 Pitch range and tone scaling

ToBI-based systems of intonation description provide symbolic labels for tagging phonologically distinctive tonal patterns and prosodic phrasing for a variety of languages. Large spoken language databases tagged using this scheme are invaluable for theoretical and computational linguistic research on topics such as pitch accent distribution, tune meaning, prosody-syntax mapping, and so on. However, for applications such as speech synthesis, where intonation must be generated from concepts (CTS) or text alone (TTS), knowing the phonological tone/phrase parse of an utterance is only the first step of the process. It is also crucial to know how these distinctive events are realized in both F0 space (vertical dimension) and time (horizontal dimension): what is often termed the *phonetic implementation*.[9]

Below I will describe the proposal put forth by Pierrehumbert and Beckman [PB88], which uses the notion of *pitch range* (defined with respect to specific prosodic domains) to provide a space within which tones are scaled. This approach has been employed in Japanese commercial text-to-speech systems, including the Bell Labs Japanese TTS and ATR's CHATR system [Spr98, CB96]. Such a model of range and tone scaling is important not only for synthesis, but is also the crucial link needed to learn from large ToBI-tagged databases. That is, it is not enough to simply observe that peak A is higher than peak B in a given discourse context. Before making such a comparison, we first need information about what phonological configurations gave rise to these peaks. Are they accented or unaccented words? Are any downstepped? Next, we need to view these contrasts within a model of pitch range and tone scaling, built from our knowledge of how tones relate to one another in the language (all else being equal). Only then can we properly interpret the observed height difference between the two peaks.

2.3.1 Scaling of the accent H* and phrasal H-

Pierrehumbert and Beckman propose that tones are scaled within an F0 space called the *pitch range*. The range is bounded on the top by a *topline* (their *h-line*), and on the bottom by a *reference line*. Both high and low tones are scaled within this range. Here I will focus only on the high tones, since they are most relevant for this thesis (but see [PB88] for a full discussion of low tone scaling in Japanese). The two high tones relevant here are the H*(+L) accent high, and the H- phrasal high. It has been observed that, all else equal, the accent high is realized at a higher F0 value than the phrasal high in Japanese (e.g. [Poser84, Sugi96, HFK86, PB88], inter alia). This is

[9] In this thesis I will only address issues of tone realization in F0 space. The reader is referred to [SP90, vSM97], inter alia for more information on the importance of F0 alignment for speech synthesis.

clear from the panels in Figure 2.1.[10] In order to account for this observation, in the model the H* tone is situated right at the topline (the highest point in the phrase), while the H- is scaled down from the topline by some fixed amount (Pierrehumbert and Beckman propose that H- is realized at 80% of the range (0.8 x topline) in their synthesis algorithm). This relation between the scaling of H* and H- is maintained even when the overall range is changed, as discussed below.

2.3.2 Downstep and the intonation phrase

In the model of pitch range proposed by Pierrehumbert and Beckman (and implemented in their synthesis algorithm), the reference line is held constant at a speaker specific value, while the topline varies depending on a number of factors. One such factor is *downstep*, a phonological process by which the range is compressed after a lexical accent. This process operates iteratively within the bounds of an intonation phrase (IP). That is, any accented accentual phrase (AP) will serve as a trigger to compress the range (defining a new topline) within which subsequent tones in the IP are scaled. This process will occur as many times as the criterion is met (within the IP), resulting in a staircase-like effect for sequences of accented APs. At the start of a new IP, the range (i.e. topline) will be 'reset' to a value which is phonologically independent of the previous specification. This manipulation of pitch range can be seen in the utterance in Figure 2.2, and is shown schematically in Figure 2.3.

Figure 2.3 shows a (schematized) sequence of four accented accentual phrases, organized into two separate intonation phrases. Due to the lexical accents in the first and second APs, the pitch range of the second AP is compressed relative to that of the first, and so is the third AP relative to the second.[11] Downstep is blocked by the presence of an intonation phrase boundary (marked by the thick vertical line), and

[10] Much of the data in previous studies involve accented words in which the accent occurs early in the accentual phrase, where it is in equivalent position to the H- (cf. the qualification of "all else equal"). However, it has been observed that words with a late accent (in which both H- and H* are apparent) often realize the accent H* peak/fall at a *lower* F0 value than the H- in the same phrase. In addition, there is some evidence which suggests that the accent H* sometimes can be scaled independently from the phrasal H-, due to variation in prominence on different morphemes in the phrase (see [VvS00]). More research is clearly needed to develop a better understanding of the influence of these factors on high tone scaling within the accentual phrase. For my purposes here, I adopt the approach described above, in which H- and H* are in a fixed relationship: H- is scaled lower than H*, all else being equal. For the short digression in Section 6.1 regarding two speakers' unaccented productions, this assumption plays a crucial role. However, for the interpretation of the majority of the data presented in this thesis, the relative relation of H- to H* is not relevant. That is, the target expressions examined in Chapters 5–7 are all monomorphemic nouns accented on the 1st or 2nd mora. Therefore, since the phrasal H- cannot be distinguished from the H* in these early-accented phrases, the scaling of H- doesn't not become a crucial issue for the data interpretation.

[11] Here, all of the APs are (early-) accented, with the accent H* tones realized right on their

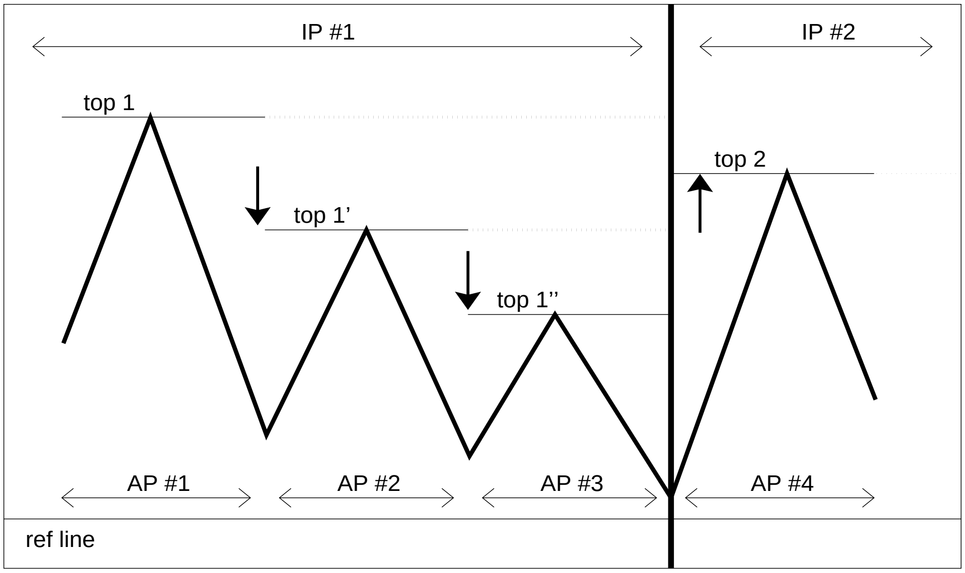

Figure 2.3: Schematic representation of changes in pitch range due to downstep and IP range 'reseting' in Japanese.

hence does not affect the range of the fourth AP. The topline of the second IP (containing this fourth AP) is 'reset' to an independent value, which in this case happens to be lower than that of the first phrase in the previous IP (AP1). Determining what (discourse) factors affect the topline specification at the start of a new intonation phrase is a topic examined by this thesis.

2.3.3 Local range variation

In addition to the pitch range of the intonation phrase, there is another factor that comes into play: the local range variation. In some cases, the pitch range of a given accentual phrase may be higher or lower than we would have predicted based on parameters determining the topline value and (fixed) degree of downstepping. For example, the range of a downstepped AP within an intonation phrase might be compressed to a greater degree than what we would expect, or the range of an AP just following an IP boundary might be expanded due to some kind of pragmatic focus. For such cases, there needs to be a mechanism by which the pitch range can be manipulated locally, at the level of the accentual phrase. Pierrehumbert and

respective toplines. Had one of the phrases been unaccented, the phrasal H- would be realized slightly below the topline, in accordance with the H- scaling discussed in Section 2.3.1 above.

Beckman adopt a hierarchical treatment of the topline in order to account for these very local changes in range. That is, in addition to the topline setting for each IP, and compression due to downstep within the IP, as described in Section 2.3.2 above, the local topline of the accentual phrase can also be changed to reflect AP subordination or prominence due to pragmatic or discourse factors. The next section discusses this in more detail.

2.4 Intonational prominence

Within a single utterance, or over a larger connected discourse, speakers may choose to highlight certain entities, or contrast them against the background discourse context. There are several acoustic and prosodic means that languages can use for highlighting certain words in a discourse, such as using larger amplitude, more extreme (peripheral) formant frequencies, increased duration, extreme/enhanced pitch movements, or prosodic phrasing, etc. Likewise, speakers can mark entities which are already salient in the discourse by using these same features, but in the reverse. Since this thesis focuses on intonation, I will briefly discuss two ways in which Japanese speakers can use intonational means to cue the salience of discourse entities.

The first way that speakers may opt to mark discourse salience is by expanding or compressing the pitch range of a phrase. Expansion of the range is a marking of *intonational prominence*, which can cue discourse ‘newness’, contrast, etc. Range compression, on the other hand, is a marking of *intonational non-prominence*, and can be used to mark discourse subordination, such as in afterthought expressions, or ‘given’ information (see the detailed discussion in Section 4.2).[12] Here, a ‘phrase’ could mean the intonation phrase (such as in marking subordination of parentheticals or afterthoughts, etc.), or it could refer to the more local accentual phrase (such as in cases of specific lexical contrast, etc.). In the hierarchical model of pitch range proposed by Pierrehumbert and Beckman, there are different mechanisms for manipulating range at these two levels, as described in Section 2.3.3 above. The range of the intonation phrase is changed by raising/lowering the topline of that phrase. This topline specification influences the realization of all subsequent tones in the IP, even in subsequent downstepped APs within the IP, since downstep is treated in the model as a fixed percent reduction of the previous topline value (and tones in downstepped phrases are scaled relative to this new topline). In contrast, the range of a single accentual phrase can be changed by raising/lowering the local topline of that phrase alone. In this case, only the tones associated with that AP (or with specific morae within that AP) are affected by the change. In running speech, speakers most likely

[12] For simplicity, I will use the general term *prominence* to refer to both prominent and non-prominent marking. However, the two contrasting terms will be used in cases where the contrast is crucial to the discussion.

mark discourse salience by manipulating range at both levels — over larger intonation phrase units, and also more locally over single accentual phrases.

A second way that Japanese speakers can mark discourse salience using intonation is by the grouping of words into prosodic units. While pitch accent is not variable in Japanese as it is in English, pitch range and prosodic grouping are variable. Both intonation phrasing and accentual phrasing are influenced by a complex interaction of phonological, syntactic, pragmatic, and discourse factors, and the accurate prediction of prosodic phrasing still presents a large challenge for modern-day speech synthesis systems. With respect to pragmatic focus, Pierrehumbert and Beckman have noted that contrastive focus on a lexical item is cued in Japanese by intonational prominence using both pitch range and prosodic phrasing [PB88]. A contrasted word will have an expanded range and will tend to start a new IP, and often the subsequent words in the utterance will be highly subordinated APs or dephrased altogether. That is, the resulting IP will start with the contrasted word marked by an expanded range, then the following words will be realized as accentual phrases in a compressed local pitch range, or may even be phrased together with the focused word into a single AP. This use of phrasing in contrastive focus is also seen in Korean, French, and several other languages (e.g. [Jun93]).

Figure 2.2 gives an example of the effect of focus on intonation phrasing. Here, the speaker is instructing the listener on where exactly to put the piece representing the window in a housebuilding task. She has already laid down the triangular roof (*sa'nkaku no ya'ne*), and instructs the listener to put the window in the **middle** (*mannaka ni*) of the roof. The word *mannaka* is made intonationally prominent by the start of a new intonation phrase. Note that this lexical item is unaccented. This provides a good example of how accentuation and intonational prominence are separate issues in Japanese (quite unlike in English). Since accent is a lexically distinctive property in Japanese, it is not an available means to cue prominence. Instead, pitch range and phrasing can achieve this goal. In the remainder of this thesis, I will focus the discussion and data analysis on the role of pitch range variation as a means of indicating intonational prominence on discourse entities. I will defer detailed examination of the role of prosodic phrasing as a prominence marker in discourse for future studies.

Figure 2.4 shows how the range variations could be interpreted in the example utterance given in Figure 2.2. I emphasize *could* here because there is no way to determine (based on just one utterance) what the separate contributions of local AP range variations and IP range variations are. A more controlled experiment involving sets of minimal pairs would be needed in order to tease apart the two.

The hypothetical pitch range schematization in this figure shows the second AP in IP1 downstepped relative to the first AP. The local toplines of both APs coincide with the original (top 1) and downstepped (top 1') toplines in the IP, and I have not opted to introduce any local AP range variation here. The accent H* tones in both

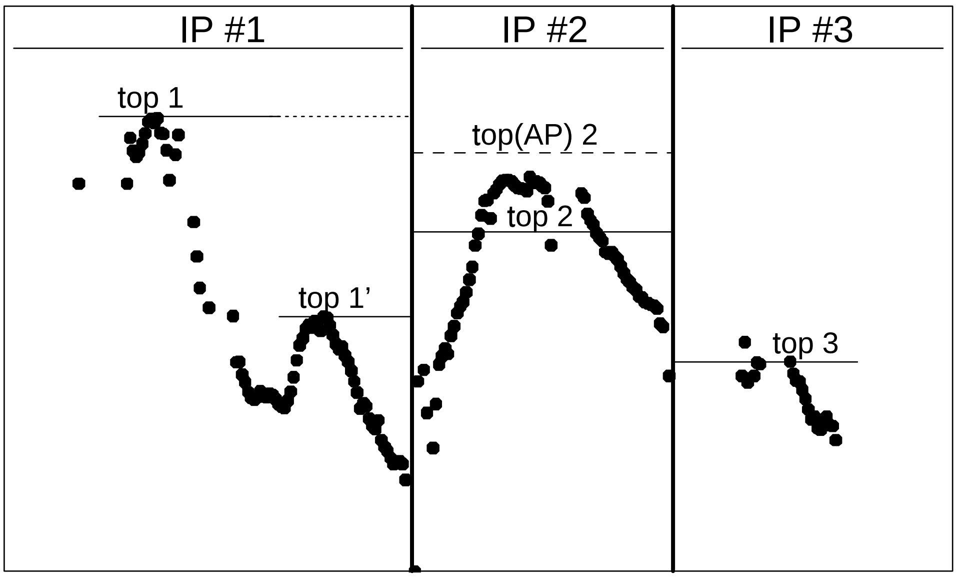

Figure 2.4: Hypothetical analysis of pitch range variation in the utterance shown in Figure 2.2 above.

APs are scaled right on their respective toplines.[13] Following this first IP, there is an intonation phrase break which functions as a prominence marking of the focused *mannaka ni*. The topline of the new phrase (top 2) is independent of the value of the previous topline (top 1). In addition, I have added an adjusted AP topline (top(AP) 2) in the figure, to show that the local range of the AP *mannaka ni* is expanded due to the intonational prominence of this AP.[14] Since this is an unaccented phrase, the H- phrasal tone is scaled down slightly from the adjusted AP topline (if it were an

[13] It is important to reiterate that the choice of overall pitch range of an IP is an indicator of intonational prominence, as is any variation of the local AP topline. However, the reduction of pitch range due to downstep is a purely phonological (automatic) effect, and should not be considered a marking of intonational non-prominence. In the same way, the scaling of H- relative to H* is governed by the tone scaling model described in Section 2.3.1, and is not reflective of prominence differences (at least not for the model assumed here, but see [VvS00] for counter-arguments).

[14] Note that this AP *mannaka ni* is also a single IP as well. It is thus impossible to tell by only this example whether the range of the phrase is due to the IP specification alone or to an additional local AP adjustment. I have opted to make the IP topline lower than the observed peak, reflecting the fact that sentence-medial IPs often have a reduced range in relation to sentence-initial IPs (for example, see the implementation in the Japanese SRS speech synthesis system [BHF83]). This topline is then adjusted to reflect the local prominence of the focused *mannaka ni*.

H* it would be realized at the topline). The last phrase containing the verb has a very reduced IP range, typical of many sentence-final predicates in Tokyo Japanese.

It should be clear from the example in Figure 2.4, and from the discussion in the preceding sections, that an analysis of peak height variation in a database cannot be undertaken without knowledge of what phonological structures give rise to the peaks. In order to learn from the vast amounts of data available in various spoken language corpora, one must (1) first tag the corpora using a scheme which marks the relevant phonological contrasts in the language's intonation system (i.e. ToBI), then (2) interpret these symbolic labels using a model of pitch range and tone scaling. With the phonological tags and interpretive model in place, databases can then be used to investigate which (discourse and other) factors are relevant for determining intonational prominence marking on referring expressions in discourse.

In the next chapter I will introduce an approach to discourse analysis that has been used to tag spoken discourse databases, and in Chapter 4 I will discuss how these discourse structures have been related to intonation variation in English, Japanese, and other languages. I will then turn to an empirical investigation of the discourse factors that can influence the realization of intonational prominence in spoken Japanese discourses.

CHAPTER 3
DISCOURSE ORGANIZATION AND ATTENTIONAL SALIENCE

3.1 Structures in discourse

A spoken utterance is not just a simple string of words or sounds. Linguists speak of the syntactic structure and the prosodic structure of an utterance, which represent the way in which words and sounds are grouped, and the hierarchical relations these groups have with respect to one another. In a similar way, a discourse is not just an unstructured string of utterances. The individual utterances of a discourse are also grouped into higher-level units, reflecting the intentions of the discourse participants. These units serve as domains by which relations between referring expressions (e.g. the relations between the definite noun phrase *the mango* or the pronoun *it*, and their antecedents) can be systematically interpreted.

Spoken language databases containing connected speech are rich not only in intonational variation, as outlined in Chapter 2, but in discourse structure variation as well. Such databases can provide the wealth of data necessary for linguistic research on numerous topics, including the discourse-intonation interface, which is the focus of this thesis. However, database analyses are impossible without a systematic tagging scheme to mark discourse structuring. One theory that has been put forth in an attempt to concretely define the nature of such structure is that of Grosz and Sidner [GS86]. This approach has been widely accepted among computational linguists working on natural language generation and understanding, and it has also been used to tag various spoken language corpora (e.g. [GH92, PL93, HN96, Naka97b]).

In the following sections of this chapter, I will briefly outline the theory of discourse organization proposed by Grosz and colleagues [GS86, GJW83, GJW95]. The approach describes a discourse in terms of three components, each of which I will discuss in turn. The *linguistic structure* represents the grouping of individual utterances into higher-level discourse units, and the *intentional structure* characterizes what the speaker intends by producing this grouping.[1] Finally, the *attentional state* models the 'accessibility' of entities to the discourse participants at any given point in the

[1] Since this thesis mainly concerns spoken language production and comprehension, I will use the term *speaker* to refer to the 'discourse producer' and *listener* to refer to the 'discourse perceiver'. However, the principles of discourse structuring apply to written discourses as well, and thus *writer* or *reader* can just as easily be substituted.

discourse. The attentional state is built on the basis of the linguistic structure and intentions, and is what constrains the interpretation of referring expressions.

3.2 Linguistic structure and intentions

Like the syntactic and prosodic structures of an utterance, which represent the linguistic organization of its words and sounds, the *linguistic structure* of a discourse represents the organization of utterances into higher-level units in connected speech (or text). Grosz and Sidner [GS86] propose that utterances are grouped into cohesive units known as *discourse segments* (DS), which serve as the building blocks that make up the entire discourse. Discourse segment boundaries are often marked by linguistic means such as specific lexical items known as *cue phrases* (e.g. *so*, *next*, *finally*, etc.) [HL87], shifts in tense, or by systematic intonational variation (see the extended discussion in Chapter 4).

According to Grosz and Sidner's approach, utterances which are grouped into a single discourse segment share a common property: they all contribute to the overall *purpose* or *intention* that a speaker has for producing that particular segment. The purposes of the segments (*discourse segment purposes* or 'DSPs') then contribute to the overall purpose of the discourse (the *discourse purpose* or 'DP'). To put it in other words, a speaker generally has a reason for initiating/producing a discourse. For example, the DP of the discourse in Figure 3.1 is to instruct a new cook how to make a Hawaiian-style breakfast. The speaker can break this instruction down into smaller 'chunks' for the new cook to follow, such as instructing how to chop up the macademia nuts, how to pit and cut the mango, how to drain the pineapple, how to assemble the dish, etc. (these sub-purposes are the DSPs). The individual utterances in the speaker's discourse contribute to the DSPs of the segments to which they belong, which in turn contribute to the overall DP.

Intentions play a major role in Grosz and Sidner's theory of discourse organization. They provide the main motivation for the grouping of utterances into discourse segments. That is, the grouping represented in the linguistic structure is dependent upon the organization of purposes in the *intentional structure*. In addition, the relations that hold among segments in the linguistic structure are based on the relations of their respective intentions in the intentional structure. Grosz and Sidner describe two ways in which segments may be related: by *dominance* or by *satisfaction-precedence*. In the case of dominance, the DSP of DS_A contributes in some way to the DSP of DS_B. For example, the DSP which instructs the cook how to remove the mango pit (DS_A) contributes to the purpose of the higher-level DSP of preparing the mango (DS_B). This dominance relation is represented hierarchically by the *embedding* of discourse segments: DS_A (pitting mango) is embedded relative to DS_B (preparing mango). In the satisfaction-precedence relation, on the other hand, DSP_A must be satisfied before DSP_B. For example, the DSP which instructs how to prepare the pineapple must be satisfied before the DSP instructing how to assemble the ingredients. This can be thought of as a *sister* relation in the hierarchical structure. In this way, using intention-driven discourse segmentation, and intention-based relations between segments, a hierarchical representation of a discourse can be formed. I will refer to this structure by the general term *discourse structure*, in order to abstract away from the specific linguistic and intentional structures which both play integral and inseparable roles in defining this structure.

3.2.1 Annotating discourse structure

Research on linguistic cues to discourse structure can be greatly facilitated by large databases systematically tagged with labels marking discourse segmentation. To this end, Nakatani and colleagues have developed a segmentation scheme based on Grosz and Sidner's model [NGAH95]. Their guidelines train labelers to identify discourse segments and their purposes, and to judge the hierarchical relations that hold among the segments. The scheme uses the `WHY?` label to tag the speaker's intentions, resulting in annotations such as that shown in Figure 3.1.[2]

Here, the labeler has identified four main sub-purposes (DSP1, DSP2, DSP4, DSP5) which contribute to the overall discourse purpose. DSP3 (how to remove the mango pit) is a sub-purpose of DSP2 (how to prepare the mango), and as such DS3 is represented as an embedded segment. DS2 is resumed after the end of this embedded segment (no additional `WHY?` label is needed for resumed segments).

In the case of instruction or task-based discourses such as this, the discourse segmentation is relatively straightforward. However, even though the main 'chunking' of the discourse is apparent, the granularity of segmentation may vary among labelers. For example, the labeler could have opted to introduce yet another super-segment containing DS1–4, which would have as its purpose to instruct the cook how to prepare the ingredients. This super-segment would then be a sister to DS5 (instruct how to assemble the ingredients). This would not affect the DS boundary locations per se, but would change the hierarchical organization. Another possibility is that the labeler could have opted to break one of the DS down further into multiple sub-purposes, such as is in DS2: the last utterance (10) could be an embedded DS in and of itself (instruct how to prevent mangos from becoming discolored). This would change both the hierarchical structure and the DS boundary locations.

There is often no one fixed segmentation for a given discourse. Indeed, it is impossible to know exactly what the speaker intended by producing a discourse, so discourse segmentation in practice amounts to a 'best guess'. However, power is in numbers,

[2] This recipe is a translation of one of the Japanese discourses constructed for the read speech database outlined in Chapter 5 (see Appendix A, Figure A.9). It describes the author's rendition of a dish served at the Volcano House Bed and Breakfast (Volcano Village, The Big Island, Hawaii). Please do not attempt to remove a mango pit like this at home — it is quite impossible.

DP: WHY? instruct new cook how to prepare a Hawaiian-style breakfast

DSP1: WHY? instruct how to prepare the macademia nuts

1 Since macademia nuts are a product of Hawaii, they are a perfect ingredient for today's breakfast menu.

2 Today we'll use crushed nuts, but even if you can't get your hands on the crushed ones, you can easily crush them by putting them in a paper bag and hitting them with something hard.

DSP2: WHY? instruct how to prepare the mango

3 The sweetness of mangos goes well with nuts, so we'll choose a ripe one and start the preparation.

4 Wash the mango clean, and carefully peel off the supermarket seal, if there is one, so that it doesn't leave a mark.

5 Cut it down the middle lengthwise leaving the skin on, and remove the pit.

> **DSP3: WHY? instruct how to remove the pit**
>
> 6 Mangos have big pits in the center like peaches, and there is a knack to removing them.
>
> 7 Firmly hold the mango skin-side down, and with a small knife, cut around the pit like you are digging.
>
> 8 If you do this, you'll be able to remove the pit easily.

9 We'll use the mango we just cut lengthwise in half, leaving the skin on, as a bowl to hold the other ingredients in today's dish.

10 Squeeze some lime on the surface so it doesn't discolor.

DSP4: WHY? instruct how to prepare the pineapple

11 We'll go on to the next ingredient.

12 Wash off the cutting board, and cut the pineapples into small pieces.

13 If you've got fresh Hawaiian pineapple, that is best, but if you don't have any, canned pineapple will do.

14 However, let's make sure to drain it before using.

DSP5: WHY? instruct how to assemble the ingredients

15 Finally, we'll assemble the ingredients at last.

16 Put some pineapple into the 'mango dish' we just made, then generously sprinkle the crushed macademia nuts on top.

17 You can put more lime on it if you like, to enjoy a fresh taste.

Figure 3.1: Discourse annotation of the Hawaiian breakfast discourse using the `WHY?` tagging scheme developed by Nakatani et al. [NGAH95].

and if multiple labelers agree on a given segmentation, this is the best approximation we have to the speaker's actual intention. Therefore, studies examining linguistic cues to discourse structuring in large databases have used consensus labels or majority labeler agreement to determine the structure (e.g. [GH92, PL93, HN96, Naka97b]).

3.3 Global attentional salience

The third component which Grosz and Sidner describe as being important in discourse production and understanding is the *attentional state*. In their words, "the attentional state ... furnishes the means for actually using the information in the other two [linguistic and intentional] structures in generating and interpreting individual utterances" [GS86, p. 177]. That is, the attentional state refers directly to the discourse structure (which I have defined above as the intention-based hierarchical segmentation), in order to determine which entities are 'accessible' or 'salient' to the discourse participants at any given point in time. As the discourse and the intentions driving the discourse evolve over time, so does the model of the attentional state, and the entities which are salient within this model also dynamically change with time.

3.3.1 Focusing structure

Grosz and Sidner model the attentional state using a *focus stack*: discourse segments are represented by spaces on the stack, and entities contained within the DS are situated within these focus spaces. When a discourse is initiated, a focus space is pushed onto the empty stack, representing the first discourse segment and its purpose. Elements (entities, events, propositions, etc.) contained within that DS are added to the stack as they are mentioned. Once an entity is added to the stack, it is said to be *accessible*, *salient*, or in *global focus*.[3] Salience has associated with it many

[3] This use of *focus* should not be confused with terms such as 'narrow focus' or 'intonational focus', often seen in the intonation literature. Grosz and Sidner's use of *focus* is orthogonal: an entity in global focus is attentionally *salient* in the discourse, that is, it is available to be talked about or referred to using a definite referring expression (regardless of any effects of intonational focus). To confuse matters more, both *focus* and *salience* have been used in the intonation literature to refer to intonational *prominence*. For example, in the utterance 'LEGUMES are a good source of vitamins', it is possible to describe the function of placing the nuclear accent (or any accent, for that matter) on 'legumes' as making the word intonationally *prominent* or *salient* to the listener. However, it is important to note that this use of *salient* is different from the use in descriptions of discourse entities. (I will return to describe the relation of these two types of salience in the remaining chapters of this thesis.) Ladd summarizes this important distinction in terminology by stating:

> "... the term *focus* is used in two essentially incompatible ways in the recent literature. In the tradition that begins with Grosz and Sidner (1986), a discourse entity is said to

privileges: for example, a salient entity may be referred to using a definite referring expression (e.g. if the speaker introduces the entity *a mango* into the space, it can subsequently be referred to as *the mango* or even *it*, so long as it remains salient). A salient entity also can be marked by intonational non-prominence (see extended discussion in Chapters 4–6). This *global* attentional salience is distinguished from *local* attentional salience in the model, as I will describe in Section 3.4 below.

On the completion of a discourse segment (when the DSP has been achieved), the corresponding focus space is popped from the stack, and its entities are no longer considered to be salient to the discourse participants. A new focus space can be then pushed onto the stack, corresponding to the next DSP/DS. The satisfaction-precedence relation in the intentional structure (see Section 3.2) is realized in the attentional state by a pop of one space followed by the subsequent push of the next. The dominance relation is realized by the push of a space *without* the pop of the previous space. That is, an embedded segment (DS_A) is represented by the push of a new space onto the stack, on top of the space corresponding to the embedding segment (DS_B). In such a case, the entities in the current space (DS_A) are said to be immediately accessible/salient (the first pick for antecedents of definite referring expressions), while those entities in the space further down on the stack (DS_B) are also thought to be salient, albeit less so. Figure 3.2 gives an example of how these pushes and pops, which update the attentional state at every turn, can be represented schematically for the cooking discourse given in Figure 3.1.

The figure shows only twelve of the many frames representing the changes to the attentional state over the course of this discourse. The frames are not sequential, but are highlights of certain key events relevant to the explanation above. Frame 1 shows a push of the space representing the whole discourse and its purpose (FS0) onto the empty focus stack.[4] There are no utterances directly related to this purpose. Frame 2 shows the push of FS1, which sits atop FS0 (since there was no intervening pop of FS0). Frame 3 shows entities from the first sentence of DS1 being added to the current empty space (FS1), and Frame 4 shows more entities being added from sentence 2.[5] When entities such as *Hawaii* or *product* are newly added to the space

be 'in focus' if it is the current topic of conversation, that is, if it is the most salient or activated in the speakers' awareness. Such entities are 'given' rather than 'new', and as such are likely to be referred to with *unaccented* expressions in English. This usage of the term contrasts with the older usage ... In this usage, focus attaches to the most informative parts of the sentence, which are accordingly likely to be pronounced *accented* in English." ([Ladd96, p. 294-295, footnote3])

[4] In this figure, the focus spaces are referred to by FS1, FS2, etc. These correspond to spaces on the stack that are open while DS1, DS2, etc. are being processed.

[5] Here, for illustration I am using the sentence as the unit to update the entities in the attentional

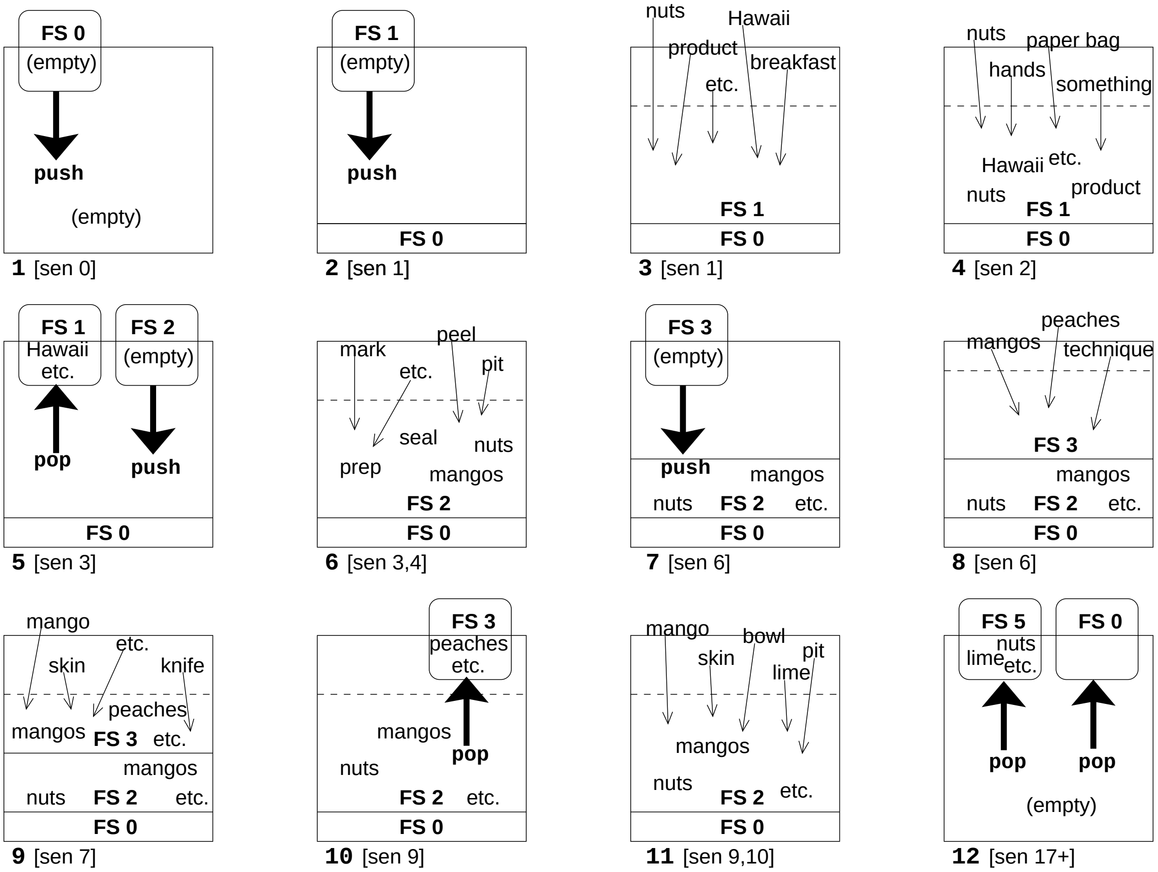

Figure 3.2: Schematic representation of dynamic changes in the attentional state of the Hawaiian breakfast discourse.

(Frame 3), they are not considered globally salient yet. It is not until they are already in the space (Frame 4) that they are considered globally salient (see Section 4.2.3 in the next chapter for intonational marking of this salience contrast). Frame 5 shows the pop of FS1 and subsequent push of FS2, representing the satisfaction-precedence relation between the DSPs in the intentional structure. Frame 6 adds entities to FS2, then Frame 7 pushes FS3 onto the stack. This push (without a pop of FS2) represents an embedding (dominance relationship among DSPs). Entities are added to FS3 in Frames 8 and 9, and are considered immediately accessible/salient, while those in FS2 further down on the stack are still salient but to a lesser degree than those in FS3. FS3 is then popped from the stack at the end of the embedded segment in Frame 10. At this point, DS2 is resumed (FS2 is reopened), and the entities in FS2 become immediately accessible again.[6] Frame 11 shows more entities being added to FS2, and finally Frame 12 skips to the end of the discourse, at which point both FS5 (the last DS in the instruction) and FS0 are popped, leaving an empty focus stack.

These push/pop operations, which mirror the intentional structure of the discourse, result in a dynamic model of attentional state which evolves over the course of the discourse. Entities are constantly going into and coming out of global focus, and it is these dynamic changes in salience that helps drive the complex interpretation (and realization) of referring expressions encountered throughout the discourse.

3.4 Local attentional salience

In addition to global attentional salience, there is another aspect of the attentional state which further constrains the interpretation and realization of referring expressions: the *local* attentional salience. A distinction is made in this approach to discourse modeling between the global versus local coherence of segments, and the extent to which each level plays a role (via the attentional state) in constraining the interpretation and realization of referring expressions. A discourse segment is *globally coherent* in that its purpose, and the propositions expressed within the DS, contribute to the overall purpose and realization of the discourse. Discourse segments are coordinated by the speaker to achieve this goal. Attentional salience of discourse entities at this level is modeled by the workings of the focus stack, as described above in Section 3.3. In contrast, the *local coherence* of a discourse segment describes the relation that the utterances within the segment (i.e. within a single focus space) have with respect to their neighboring utterances. Properties of the attentional state at

state. It is possible that a more appropriate unit would be the clause, the intonation phrase, or something else that combines syntactic and prosodic junctural functions.

[6] But see [Walk98] for a discussion of the *inaccessibility* of entities in a resumed DS after a long embedded segment.

this more local level serve to further constrain the interpretation and realization of referring expressions. Grosz and colleagues [GJW83, GJW95] model local coherence within discourse segments in terms of discourse *centering* processes.

3.4.1 Centering Theory

Discourse centering, or *Centering Theory*, models the dynamics of attentional salience at the local level. That is, it describes the relative salience of entities *within* a single utterance, and provides mechanisms for predicting which of these entities is most salient — i.e., what the utterance is most centrally 'about'. This in turn constrains the form that referring expressions may take: an entity which is most salient at this level, in *local focus*, generally is realized using a pronoun in English. In addition, Centering makes predictions about what types of utterance sequences are preferred over others in discourse comprehension/processing: sequences in which the local focus is maintained over a stretch of speech are preferred over those in which the focus is constantly changing.

This approach has received much attention recently, and has been tested for several languages, including English, Hebrew, Italian, Japanese, and Turkish (see the studies included in [WJP98a]). In the following sections I will introduce the basic notions of Centering, then discuss how it has been applied specifically to Japanese.

Example data

This section presents six example (mini-)discourses which highlight the phenomena that Centering Theory aims to describe. The questions that are enumerated below regarding the data will be answered in the following sections which outline the workings of the theory.

(1) a. Susan gave out candy for Halloween.
    b. She had bought the candy at Revco on sale.

In (1a) three discourse entities are introduced into the global focus space (*Susan, candy, Halloween*). In (1b) two of them are referred to again (*Susan, candy*) and another is introduced (*Revco*). In (1b), the global attentional salience of both *Susan* and *candy* (that is, they are already located in the focus space when (1b) is uttered) licenses the use of the definite referring expressions *she* and *the candy* to refer to these entities. However, why did the speaker choose a pronoun (*she*) over a full definite noun phrase (*Susan*) in (1b)?

(2) a. Susan gave Betsy a pet hamster.
    b. She reminded her that such hamsters were quite shy.
    [GJW95, Example 6]

In (2b) its possible that either pronoun could refer to either person mentioned in (2a), since both are female 3rd person referents. Why then do native listeners prefer to interpret (2b) as meaning 'Susan reminded Betsy ...'? As Grosz et al. note, this preference becomes clear when a subsequent utterance (c) is added, as in the following minimally different discourses in (3) and (4):

(3) a. Susan gave Betsy a pet hamster.
b. She reminded her that such hamsters were quite shy.
c. She asked Betsy whether she liked the gift.
[GJW95, Example 7]

(4) a. Susan gave Betsy a pet hamster.
b. She reminded her that such hamsters were quite shy.
c. She told Susan that she really liked the gift.
[GJW95, Example 10]

Why is discourse (3) much easier to understand than discourse (4)? And finally, why does discourse (5) feel much more coherent than discourse (6), even though they both have (nearly) the same content?

(5) John went to his favorite music store to buy a piano.
He had frequented the store for many years.
He was excited that he could finally buy a piano.
He arrived just as the store was closing for the day.
[GJW95, Example 1]

(6) John went to his favorite music store to buy a piano.
It was a store John had frequented for many years.
He was excited that he could finally buy a piano.
It was closing just as John arrived.
[GJW95, Example 2]

Based on these examples, and many others like them, Grosz et al. define a set of Centering structures, relations, and principles that determine (a) when pronouns can be used, (b) which entity in the local discourse context a pronoun will refer to, and (c) what relations among utterances result in a 'coherent' discourse. The following sections outline the core notions of Centering Theory.

Centering structures

The notion of a *Center* attempts to capture the fact that most utterances seem to 'center' around a particular entity: that is, the most salient entity in the discourse at a given point in time. For example, both utterances in example discourse (1) center around Susan, as do the utterances in (2) and (3). Discourse (5) is about John. In each of the utterances in (5), John seems to be in the center of attention — he is what the utterances are 'about'. Grosz et al. use the term *backward-looking Center* (Cb) (or more generally, *the Center*) to refer to this entity that the utterance is centrally 'about'. The Cb, being the most salient entity in the local discourse context, is typically realized by a reduced linguistic form, such as a pronoun in English. As the name suggests, the Cb also serves to link the current utterance (U_n) with the previous one (U_{n-1}). That is, of the entities contained in U_{n-1}, one of them is chosen to be the Cb in the following utterance (U_n). The designation of an entity in U_{n-1} to be the Cb of the next utterance is not random, but is determined by the salience ranking of all entities found in that utterance (U_{n-1}). These entities are called the *forward-looking Centers* (Cfs), and are potential candidates to become the Cb in the next utterance.

Forward-looking centers are ranked according to their salience in the utterance. The ranking most often used in the literature for English (and for other languages as well [WJP98a]) is according to grammatical role:[7] [8]

subjects > objects > other

The claim is that grammatical subjects are more salient in the local discourse than grammatical objects or other entities. This has been noted in the literature (e.g. [Chafe76, Prince92]), and has been shown by psycholinguistic experimentation [GGG93]. Among the entities found in the Cf list, the highest ranked member (the subject, if there is one) in the current utterance U_n is termed the *preferred Center* (Cp), and it is the most-likely candidate for being the Cb of the following utterance U_{n+1}. In this way, the Cp is a prediction about what U_{n+1} will be 'about'.

Centering transitions

As outlined above, Centering Theory defines for each utterance a set of ranked forward-looking Centers (the highest ranked of which is the preferred Center), and one backward-looking Center indicating which of the Cfs in the previous utterance the current utterance is most centrally 'about'. Transition relations between adjacent utterances can be defined by the relationship among these Centering structures:

[7] Note here that, in a majority of the Centering literature, for practical purposes only discourse entities (elements referred to by noun phrases (NPs)) are considered to be possible Centers, though most works do acknowledge that events and even whole propositions could also serve as Centers. However, since the behavior of nouns with respect to Centering is the most clearly understood at this point in time, I will limit the discussion and analysis in this thesis to NPs only.

[8] Also see [WJP98a] for alternative proposals for ranking Cfs.

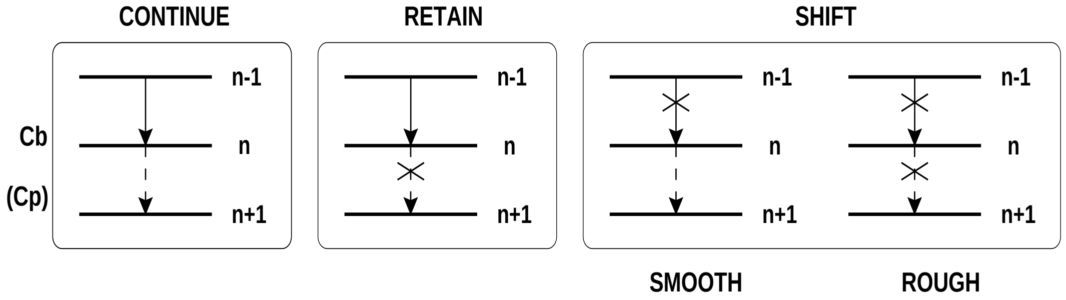

Figure 3.3: Schematic representation of Centering transition types currently recognized by Centering Theory.

specifically, the backward-looking Centers (Cbs) of utterances U_{n-1} and U_n, and the preferred Center (Cp) of U_n. A speaker can choose to *continue* talking about the same entity over a string of utterances, or she can *shift* the local salience to another entity in the discourse. Figure 3.3 outlines the Center transition types that are currently recognized in Centering Theory [WJP98a].

The three Center transition types are *continue*, *retain*, and *shift*, corresponding to the boxes in the figure. In addition, shift transitions can be further divided into *smooth shift* and *rough shift.* Center *continuation* occurs when the entity which is Cb(U_{n-1}) is the same as Cb(U_n), and it is also the Cp of U_n (indicating that it will become the Cb of U_{n+1}, if it is realized there). That is, a continuation occurs when the speaker has been talking about an entity X (i.e. it is in local focus), and will continue to talk about it. Center *retaining* is similar, in that Cb(U_{n-1}) is the same as Cb(U_n). However, in this case, Cb(U_n) is not the Cp (for example, the Cb may be in the less salient object position), and thus that entity will not continue to be Cb in U_{n+1}. Grosz, et al. suggest that retaining may be a means by which speakers can "produce a smooth transition to a new center" [GJW95, p. 215]. Finally, a Center *shift* is defined by the fact that Cb(U_{n-1}) is the not the same as Cb(U_n). That is, shifts occur when the speaker has been talking about an entity X, but then changes the local center of attention to another entity Y. The distinction between smooth vs. rough shifting is based on the identity of Cp(U_n). If Cb(U_n) is the same as Cp(U_n), then the transition is a *smooth shift.* This indicates that the speaker was talking about an entity X, but has shifted the local focus to an entity Y, and intends to continue talking about Y. On the other hand, if the Cb(U_n) is not the same as Cp(U_n) (e.g. it is realized in a less salient grammatical position), then the transition is defined as a *rough shift.* In this case, the speaker shifts the focus from X to Y, and intends to shift again to Z. Table 3.1 (from [WJP98b]) summarizes the differences among the four transition types, in terms of the status of the Cb and Cp.

CONTINUE	$Cb_n = Cb_{n-1}$ (or $Cb_{n-1} = \emptyset$)	and	$Cb_n = Cp_n$
RETAIN	$Cb_n = Cb_{n-1}$ (or $Cb_{n-1} = \emptyset$)	and	$Cb_n \neq Cp_n$
SMOOTH-SHIFT	$Cb_n \neq Cb_{n-1}$	and	$Cb_n = Cp_n$
ROUGH-SHIFT	$Cb_n \neq Cb_{n-1}$	and	$Cb_n \neq Cp_n$

Table 3.1: Summary of Cb and Cp relations in the different Center transition types.

Example (7) shows the Centering structures and transitions for the discourse in (3) above.[9]

(7) a. Susan gave Betsy a pet hamster.
[**Cb**=$\emptyset$*;* **Cf**=*Susan* > *hamster* > *Betsy;* **Cp**=*Susan*] NULL
b. She reminded her that such hamsters were quite shy.
[**Cb**=*Susan;* **Cf**=*Susan* > *Betsy* > *hamsters;* **Cp**=*Susan*] CONTINUE
c. She asked Betsy whether she liked the gift.
[**Cb**=*Susan;* **Cf**=*Susan* > *Betsy* > *gift;* **Cp**=*Susan*] CONTINUE

In this discourse, *Susan* is introduced as a subject NP into the focus space in (7a), and by virtue of its syntactic position, this entity is the highest ranked in the Cf list (Cp). This Cp becomes the Cb of (7b) by the Centering principles (below), and again it receives Cp status because of its subject-hood. It continues to be the Cb of the subsequent utterance (7c). The transitions between (7a) & (7b), and (7b) & (7c) are both characterized as *continues.*

Centering transitions describe the way in which adjacent utterances are linked, in terms of their locally salient entities. Along with the Centering structures Cb, Cf and Cp, transitions (especially continues) provide the means for speakers to achieve local coherence within the discourse segment.

[9] The first Center transition in this example is NULL because Centering is usually defined only *within* discourse segments. In this example, the Cb of the discourse absolute-initial utterance is undefined, since there is no preceding utterance to provide a Cf list. However, there has been question recently about whether Centering transitions can be defined within discourses but *across* DS boundaries (see the discussion in [WJP98a], and in Section 5.2.3). Also, see [WIC94] for a discussion of *Center instantiation* of topic-marked NPs in Japanese DS-initially, which I will return to in Section 3.4.2 below.

Centering principles

In this section I will reiterate the main principles of Centering Theory, summarized by [WJP98b, p. 3-4], and show how each is relevant for accounting for the data in (1)-(6) above.

- **An utterance has a set of forward-looking Centers (Cf).** As mentioned above, the entities realized in an utterance are ranked according to their local salience. Since subjects are more salient than other entities such as objects, the ranking is usually determined by grammatical role.

- **The backward-looking Center (Cb) of U_n is defined as the highest-ranked member of the Cf list (Cp) of U_{n-1}.** If this Cp is not realized in U_n, then the next highest-ranked entity on the Cf list of U_{n-1} will be realized as Cb of U_n.

- **Transitions are ordered by preference: continue > retain > smooth shift > rough shift.** The claim is that there is a psychological preference for continuing the Center in order to limit the amount of inference needed to process the discourse. That is, we 'expect' the speaker to continue talking about the most salient entity, and only need to infer differently if something in the signal forces us to do so. For example, in discourses (3) and (4), the subject *Susan* in (3/4a) is the Cp. The preference for continue, in combination with the Cp(U_{n-1})=Cb(U_n) principle (above), makes us interpret the subject Center *she* of (3/4b) as 'Susan'. That Center is then continued yet again (by the same combination of principles) to (3/4c). At the point where we encounter *she* in (3/4c), we interpret it to mean 'Susan'. This leads to the felicitous interpretation of (3). However, when encountering the full NP *Susan* in (4c), it produces a gardenpath-like effect, and reanalysis is required. The principles of Centering Theory correctly predict the infelicity of (4).

  Discourses (5) and (6) are another example of transition preference ranking. In (5), the subject *John* of the first utterance is made the Cb (and Cp) of the next utterance, and is continued as the Center throughout the discourse. In (6), the Cb constantly shifts between *John* and *the store*, and so the discourse feels much less coherent. This difference in perceived local coherence is accounted for in Centering by the ranking of transition preferences.

- **If any member of Cf(U_{n-1}) is realized as a pronoun in U_n, then the Cb(U_n) will also be realized as a pronoun.** This principle reflects the observation that the Cb of an English utterance is most often realized as a pronoun. In discourse (1), *Susan* is the highest-ranked member of the Cf list in (1a), and becomes the Cb realized by the pronoun *she* in (1b). This principle also allows for an alternative felicitous realization of (1b) as *She had bought it* *at Revco on sale*, but rules out the possibility of *#Susan had bought it at Revco on sale* as a felicitous continuation of (1a).

- **An utterance will have exactly one backward-looking Center.** In some utterances there may be a number of entities that could potentially serve as the Cb (e.g. a number of entities realized by pronouns), but only one of these is singled out and given Cb status. This is shown in examples (2)–(4). In (2b) there are two potential candidates for what the utterance is 'about' (i.e. *Susan* vs. *Betsy*). However, it is clear from the continuations of the same discourse in (3) and (4) that there is a preference for the (b) utterance to be 'about' *Susan*. This Cb is then continued on to the next utterance in (3c).

Centering Theory attempts to model the perceived local coherence of discourse segments, and the form of referring expression chosen by a speaker at any given point in the discourse. It models the *local salience* of discourse entities as they come into and go out of the focus of attention of the discourse participants. The next section discusses how the structures, transitions, and principles describing Centering in English have been adapted to describe Japanese discourse phenomena.

3.4.2 Local attentional salience in Japanese

Interpretation of zero pronouns

The Centering approach outlined above has been adopted by many researchers to describe the perceived local coherence of Japanese discourses, and to account for the distribution and interpretation of *zero pronouns* (e.g. [Kame85, Kame86, Kame88, WIC94, Naka92, NN94, TD94, Shima96, Iida98]). Zero pronouns (henceforth 'zeros') are subcategorized verbal arguments which are not realized in the surface form of an utterance. Kameyama suggests that zeros are functionally equivalent to (unaccented) pronouns in languages such as English [Kame85]. Discourse Centering processes have been proposed to constrain the interpretation of such zeros in Japanese, and in other languages including Turkish and Italian (see [WJP98a]).[10]

Example (8) gives a short Japanese discourse which contains zeros in utterances (b) and (c). In (c), the subject and object of the verb *kikimasita* 'asked' are both

[10] In Japanese the 1st person referent 'I' or 2nd person 'you' are usually realized by zeros, regardless of the discourse context. They are considered to be permanently in focus, and as such can be realized with a zero even if no overt antecedent exists. Kameyama highlights the distinction between these and other discourse referents: "1st and 2nd person references are INDEXICAL ... while 3rd person reference is usually DISCOURSE ANAPHORIC." [Kame85, p. 103, footnote 11]. Because of this discourse-independent nature of 1st and 2nd person pronouns/zeros, descriptions of Centering in Japanese (and also in English and other languages) restrict attention mainly to 3rd person discourse referents.

zero pronouns, and there are two 3rd person entities already salient in the discourse (Taroo and Ziroo), both of which would be equally appropriate for either role. This makes the interpretation of the two zeros potentially ambiguous. However, Centering correctly predicts that (c) is interpreted as 'Taroo asked Ziroo ...'. Below I will describe how Centering has been applied to Japanese to explain such examples.

(8) a. *Taroo-ga kooen-o sanpo-siteimasita.*
Taroo-SUBJ park-in walking-was
'Taroo was taking a walk in the park.'

b. *∅ Ziroo-o hunsui-no mae-de mitukemasita.*
∅-SUBJ Ziroo-OBJ fountain-POSS front-in found
'He [Taroo] found Ziroo in front of the fountain.'

c. *∅ ∅ kinoo-no siai-no kekka-o kikimasita.*
∅-SUBJ ∅-OBJ2 yesterday-POSS game-POSS score-OBJ asked
'He [Taroo] asked him [Ziroo] the score of yesterday's game.'
[WIC94, Example 2]

One of the most basic principles of Centering is that entities in U_{n-1} (Cfs) are ranked according to their local salience, and that the highest-ranked member of this set (Cp) is chosen to the the Cb of U_n. In English the ranking is done according to grammatical role. However, Japanese introduces a few language-specific grammatical phenomena that need to be incorporated in the Cf ranking. Specifically, Japanese uses morphological means (the postposition *-wa*) to mark the 'topic' or 'theme' of an utterance [Kuno73]. In addition, some verbs such as the 'giving-verbs' *kureru*, *ageru*, and so on, designate certain arguments as the locus of the speaker's empathy (e.g. [KK77]). Both topic- and empathy-marked NPs are said to be highly salient in the local discourse context, and as such are ranked high on the Cf list. Walker et al. [WIC94] have formulated the following Cf ranking for Japanese (adopted from Kameyama's Expected Center Order [Kame85]):

topic > empathy > subject > object2 > object > others

Although most agree on this ranking of local salience in Japanese, researchers differ as to what Centering principles apply to account for the correct interpretation of zeros. I will briefly describe the two main approaches here: those put forth by Kameyama [Kame85, Kame86, Kame88] and by Walker et al. (e.g. [Naka92, WIC94, Iida98]), and refer the reader to the original works for full details.

In Kameyama's account, the interpretation of zeros in (8b/c) is achieved by two separate mechanisms: *Cb-establishment* and the *Property-sharing Constraint*. Which mechanism is employed for a given sequence of utterances depends on the surface realization of the coreferent entities in those utterances. Cb-establishment is used in cases where a zero pronoun in U_n is coreferent with an entity realized by a full NP in the previous utterance (U_{n-1}). In this process, the entities in U_{n-1} are ranked according to the Expected Center Order (see ranking above), and the highest-ranked NP serves as the antecedent of the following zero, in much the same way as the Centering principles describe Cb selection (see Section 3.4.1 above). This accounts for the interpretation of the zero in (8b) as *Taroo.* The interpretation of the subject zero in (8c), on the other hand, is achieved by another mechanism. The *Property-sharing Constraint*, used in cases where zero pronouns in adjacent utterances are coreferent, dictates that "two zero pronominals that retain the same Cb in adjacent utterances should share one of the following properties (in order of descending preference): 1) Ident-SUBJECT, 2) Ident alone, 3) SUBJECT alone, 4) nonIdent-nonSUBJECT" [Kame86, p. 205].[11] This constraint links the two subject zeros in (8b/c), and gives both the interpretation *Taroo.* Thus, Kameyama's account explains the data in (8) by a combination of local salience ranking and grammatical parallelism.

In contrast, Walker et al.'s proposal accounts for the data by using the same Centering principles introduced above for English. That is, entities in a given utterance U_{n-1} are ranked according to their local discourse salience (see the language-specific ranking for Japanese above), and the highest-ranked entity in the Cf list (Cp) is chosen as the Cb of the next utterance (U_n). These principles, in combination with the preference for the continue transition, account for the interpretation of the zeros. In (8b), the zero refers to *Taroo*, which is the Cp of the previous utterance (8a), and the subject zero of (8c) also refers to *Taroo* by the same mechanism.[12] Since Walker et al. are able to account for the zero interpretation in (8b/c) using the same set of Centering mechanisms which have also been used to describe pronoun/zero interpretation in a number of other languages, for convenience I will adopt this proposal and the general Centering terminology to describe the data examined in this thesis (see Chapters 5–7).

Topic-marked NP-*wa*

Accounts of Centering in Japanese discourse segments restrict attention to the distribution and interpretation of zero pronouns as Centers (Cbs), and little mention is made of overt topic-marked entities (henceforth 'NP-*wa*') as possible Centers. However, the postposition *wa* marks an overt anaphoric expression which refers to an entity which is salient to the hearer in some way. In addition, an NP-*wa* entity is said to be in local focus, in that it is what the utterance is 'about' or the *theme* [Kuno73]. How then does this property interact with discourse Centering processes?

[11] Here, Kameyama's *Ident* refers to the empathy locus.

[12] Neither Kameyama's nor Walker et al.'s proposals directly accounts for the interpretation of the non-Cb zero in (8c), just as the realization of 'Betsy' as *her* in (2b) is also not accounted for by the English Centering principles.

Kameyama notes that while Centering in English "seems to assert that only pronouns can encode Cbs, it is unlikely. Especially, definite NP (the NP) may also encode a Cb in certain cases. Such a non-pronominal Cb-encoding is a topic for a future study" [Kame85, p. 98]. This statement can be generalized to NP-*wa* in Japanese, and in fact Kameyama also states that the overt NP which is "closely related to the Center in Japanese seems to be the TOPIC whether or not consisting of a full or pronominal NP" [Kame85, p. 130]. Moreover, Kuno claims that using a zero is equivalent to marking an entity with *wa* [Kuno72]. So what role does NP-*wa* have in the current approaches to Japanese Centering described above?

Both Kameyama and Walker et al. propose that the local salience of NP-*wa* in Japanese results in its place as the highest-ranking entity on the Cf list. This in turn is used to predict the interpretation of a subsequent zero. Walker et al. go one step further, and propose that *wa*-marking of an NP discourse segment-initially instantiates that entity as the Center (Cb) of the utterance [WIC94, p. 215]. This Cb-instantiation provides the means by which to continue the Center from a DS-initial utterance to the second utterance, as in cases such as (9).[13] However, Walker et al. do not go as far as to automatically assign Cb status to other DS-medial *wa*-marked NPs.

(9) a. *Taroo-wa Ziroo-o min'na-no mae-de tatakimasita.*
Taroo-TOP/SUBJ Ziroo-OBJ everyone-POSS front-in hit
'Taroo hit Ziroo in front of everyone.'
[**Cb**=*Taroo;* **Cf**=*Taroo* > *Ziroo* > *everyone;* **Cp**=*Taroo*] NULL

b. *Itiniti-zyuu kanzen-ni ∅ ∅ musi-simasita.*
day-throughout completely ∅-SUBJ ∅-OBJ ignored
'He [Taroo] ignored him [Ziroo] all day.'
[**Cb**=*Taroo;* **Cf**=*Taroo* > *Ziroo;* **Cp**=*Taroo*] CONTINUE
[WIC94, Example 31]

In sum, the role of NP-*wa* entities in current approaches to Japanese Centering has mainly been to aid in the interpretation of zeros in subsequent utterances. Centering makes predictions about the distribution of NP-*wa* in discourse only indirectly through predictions about the distribution of zeros: they are predicted to occur where zeros are not licensed to occur. For example, we would not expect to see a zero in a DS-initial utterance, since there would be no preceding overt NP to serve as its antecedent within the segment.[14] In this position, we might expect to find an overt NP instead, such as a *ga*-marked subject or *wa*-marked topic. The use of NP-*wa*

[13] In Walker et al.'s study, 10 of 14 subjects preferred this interpretation [WIC94, p. 215].

[14] But see [Walk98, Pass98] for discussion of cases of reduced pronominal forms DS-initially.

in DS-initial position (i.e. when the entity is first added to the current global focus space) is possible if the entity is in some way salient already. For example, NP-*wa* may be used to re-introduce into the current space an entity which was salient in another segment (i.e. in a previous focus space), or to introduce an entity which is situationally-salient to both speaker and hearer. One example of this is the use of *makademianattsu-wa* to introduce 'macademia nuts' discourse-initially in Figure 3.1: the listener is aware of the ingredient list before the speaker starts her instruction.[15] However, the exact distribution and use of overt anaphoric expressions such as NP-*wa* in Japanese is still an open research area (see e.g. [CD87, HMI87]), and Centering Theory only indirectly addresses this question.[16] In the following chapters, I will return to the discussion of NP-*wa* Centers in Japanese, and will examine the details of intonational marking on them.

[15] The first sentence of the Japanese discourse translated in Figure 3.1 reads: <u>*Makademianattsu-wa*</u> *hawai-no meibutsu de, kyô-no chôshokumenyû-ni pittari-no sozai desu.* 'Since <u>macademia nuts</u> are a product of Hawaii, they are a perfect ingredient for today's breakfast menu.' (see also Figure A.9).

[16] Craige Roberts (personal communication, July 2000) suggests that one reason an NP-*wa* may be used DS-internally is if the use of a zero pronoun would be ambiguous.

CHAPTER 4
THE INTONATION–DISCOURSE INTERFACE: A REVIEW

In Chapters 2 and 3, I outlined two approaches to the description of intonation and discourse structures, respectively. Each of these approaches has grown out from a history of research and data analysis in their respective disciplines, resulting in models of intonation and discourse which can be used to systematically tag linguistic structures found in large spoken language databases. In this chapter, I will describe research that has merged the two disciplines, in an attempt to characterize the effect that discourse structures have on intonational realization, or in reverse, the role that intonation plays in cueing discourse structures. I use the plural term 'discourse structures' here because there have been at least two separate themes of research in the intonation-discourse interface: one theme has been to examine the relation of intonation to the linguistic structure of discourse itself, and the other theme has been to examine the use of intonation in marking discourse salience, for example the traditional 'given/new' distinction. In the following sections I will give an overview of each of these lines of research in turn, and will attempt to interpret results from previous studies in terms of the intonation and discourse frameworks described in the preceding chapters.[1] I will then highlight some open research questions regarding the intonation-discourse interface in Japanese specifically, which the remainder of this thesis addresses.

4.1 Intonational cues to discourse structure

Previous studies examining the intonation-discourse mapping have found a systematic effect of discourse structure on pitch range variation.[2] The general observation

[1] There is a wealth of research on both of these aspects of the intonation-discourse interface, in a variety of languages, and I clearly could not review every such work in detail here. However, the discussion will highlight results from key studies, which will suffice to serve as a background for the experimental design and analyses presented in the following chapters. The reader is referred to the original works, and the references cited within, for additional details.

[2] Many other acoustic-prosodic cues to discourse structure have also been reported in the literature, for example pause distribution, duration, amplitude variation, melodic cues, rate effects, etc. (see e.g. [SG94, HN96], inter alia for more details). However, I will restrict discussion in this chapter to pitch range and accent, which is the main focus of this thesis.

is that pitch is raised at the beginning of discourse units, and is lowered at the end of such units. However, while studies of various languages have concluded that intonation, specifically pitch range variation, can be used to cue discourse structure, the independent definition of such structure has varied greatly.

Lehiste used the written 'paragraph' to define the discourse unit of interest [Leh75]. She found that English utterances with high F0 peaks are perceived by listeners as being paragraph-initial. Paragraph-medial and final utterances tend to have lower F0 peaks. Silverman found that manipulating the pitch range of intonation phrases in English using resynthesis can induce listeners to segment discourses with ambiguous structures differently: phrases with an expanded pitch range are likely to be judged paragraph-initial, and final-lowering of F0 can cue paragraph finality [Silv87]. Venditti studied intonation cues to written paragraph structure in Japanese, and found that discourse-initial sentences are produced in an expanded range when compared to the same sentences in discourse-final position [Ven96]. However, Venditti did not observe a robust difference between medial and final position, suggesting that Japanese uses pitch range cues primarily to mark the start of discourse units.

Other studies have defined discourse units according to 'topic' structure: stretches of speech in which the speaker is mainly discussing a single entity. Yule suggests that intonation can be used to mark the boundaries of topic units in English spontaneous speech: intonational structuring which he terms the 'paratone' [Yule80]. Swerts and Geluykens examined topic units in Dutch, defined in their housebuilding task as a stretch of speech in which a specific piece of the house is being described [SG94]. They also found that F0 is high at the beginning of such units, and gradually declines to the unit end. Venditti and Swerts used the same housebuilding task to examine intonational cues to topic structure in Japanese [VS96]. They found that the F0 height of the vowel /a/ in the sentence-final verbal affix *masu* is dependent on the location of the verb in the topic unit: *masu* in unit-initial sentences is realized in a higher range than that in medial or final utterances. This again suggests that Japanese marks discourse units using expanded range at their beginnings. However, Venditti and Swerts showed that this effect interacts with downstep in Japanese: the discourse structure effect is observed only when the verb is not downstepped by a preceding accent in the intonation phrase. In cases of downstep, the F0 range is already so compressed that any effects due to discourse position are negligible.

Topic units as defined above may in some cases be likened to intention-based discourse segments defined by Grosz and Sidner [GS86, NGAH95]. As described in Section 3.2, Grosz and Sidner's theory identifies discourse segments as those stretches of speech which contribute to a specific speaker purpose or intention. This approach to discourse structuring has been used to tag English spontaneous and read speech databases, and the relation of linguistic cues (both intonational and otherwise) to this segmentation has been studied extensively (e.g. [HP86, LH90, GH92, PL93, HN96, Naka97b]). With regard to intonation specifically, Hirschberg and colleagues have found that an increased overall F0 range is correlated with (intermediate) prosodic phrases which labelers agree to be discourse segment-initial, relative to other DS positions [GH92, HN96]. Likewise, a lower F0 range correlates with DS-final judgments. In addition, Hirschberg and Nakatani found that these effects of DS-position can be identified on-line by examining the local F0 change between one phrase and the next [HN96]. In their seminal work, Hirschberg and Pierrehumbert showed that intonational variation according to intention-based discourse structuring can improve the intelligibility of English synthesized speech [HP86]. They showed that pitch range variation can not only be used to mark DS boundaries, but also to cue the hierarchical relations (such as embeddings) among the segments. Along these same lines, Ayers has also observed that, in English, phrases which begin 'topic units' have a relatively higher pitch range than those which begin subtopics [Ayers94].

In sum, both English and Japanese (and other languages as well) use pitch range manipulations to cue discourse structuring by marking the edges of discourse units. An expanded range marks the beginnings, and a compressed range marks the ends of such units. However, while there are many studies citing the relation of intonation to discourse structure, the exact nature of this structure is not often independently and systematically defined. It is encouraging that studies using Grosz and Sidner's intention-based approach to discourse segmentation have observed similar effects as in previous studies. Given the effects of F0 raising in initial position reported for Japanese written paragraphs and topic units [Ven96, VS96], I hypothesize that the same raising effect will also be found for initial phrases defined using intention-based discourse segmentation.

4.2 Intonational cues to discourse salience

A second line of research in the intonation-discourse interface has been to determine the intonational correlates of discourse salience in various languages. The notion of 'discourse salience' has been defined in nearly as many ways as there are studies on the subject, and I will describe a few approaches to defining this notion in the following sections.

4.2.1 'Given' vs. 'new' information

The *given* vs. *new* dichotomy often cited in the discourse literature was defined by Halliday directly in terms of the speaker's choice of intonational form [Hal67]. For Halliday, *new information* is focal information which "the speaker presents ... as not being recoverable from the preceding discourse" (regardless of whether or not it had been mentioned before) [Hal67, p. 204]. New information is marked by intonational prominence, by a 'tonic' or 'nuclear' pitch accent in English.[3] *Given information*, on

[3] The English ToBI approach to intonation description defines a *nuclear* (as opposed to *prenuclear*) pitch accent as the last accent in the intermediate phrase. This typically corresponds to the most

the other hand, is considered by Halliday as that which is not new, and is marked by the intonational non-prominence of the lower-pitched prenuclear portion preceding the nuclear accent.

While Halliday's claim is that the given/new distinction is defined solely by the speaker's choice of intonational grouping and prominence, subsequent studies have attempted to relate the intonational phenomena to independent text-based characterizations of given vs. new information. For example, Brown used Prince's [Prince81] taxonomy of discourse givenness to describe variations in intonational prominence in English task-oriented speech [Brown83]. Brown found that speakers tend to place intonational prominence on new information (Prince's *brand-new* and *inferred*), while marking given information (Prince's *situationally* or *textually-evoked*) with intonational non-prominence.[4] In English, the means for indicating intonational prominence is via *pitch accent.* That is, new information tends to be *accented* and given information tends to be *unaccented* or *deaccented.*[5]

In Japanese, in contrast, the notion of accent cannot be used to mark discourse salience. As described in Section 2.2 above, the accented/unaccented distinction in Japanese is tied to the lexical item itself, and not to any discourse function. However, the language does have different means, namely pitch range variation and prosodic

prominent accent or the 'sentence stress' (e.g. [CH68, Pierre80, BP86], inter alia). Here, I am interpreting Halliday's observations about tonic/pretonic portions of the tone group in terms of this nuclear/prenuclear distinction.

[4] These categories are defined by Prince as follows: *brand-new* entities are not previously known to the hearer, while *unused* entities are those which are known to the hearer, but are not previously part of the discourse model. *Textually-evoked* entities are those already in the discourse model by means of previous mention, *situationally-evoked* entities are those which "the hearer [knows] to evoke ... all by himself, for situational reasons" (e.g. the pronoun *you*), and *inferable* entities are those which the speaker assumes the hearer can infer from other evoked discourse entities (e.g. *the driver* is inferable from *the bus*) [Prince81, pp. 235-236].

[5] However, it should be noted that Halliday's and Brown's observations of intonational prominence/non-prominence do not directly equate with the *accented/unaccented* distinction known in ToBI-like systems. That is, while both Brown and Halliday associate new information with the prominent nuclear accent, nothing is said about possible other prenuclear accents, which are also said to be prominent in relation to unaccented words (e.g. [Pierre80, Ayers96]). In addition, Brown and Halliday characterize given information by a "lack of phonological prominence [that] yields a syllable close to the baseline with little, if any pitch movement" [Brown83, p. 73]. This characterization of non-prominence could be interpreted in a ToBI-like framework as *unaccented* material, though not all unaccented words have an F0 near the speaker's baseline, and L* accents also "yield a syllable close to the baseline". However, since in many cases these notions of intonational prominence/non-prominence do coincide with the *accented/unaccented* distinction, and because many subsequent studies have used them as such (e.g. [Terk84, TH94, Naka97a], among others), I will adopt the ToBI *accented/unaccented* terminology here to describe the English phenomena.

phrasing, that can be used to cue intonational prominence (for example, in cases of contrastive focus), as outlined in Section 2.4. As mentioned earlier, in this thesis I will focus only on the use of pitch range variation in cueing prominence, and defer further investigation of phrasing and prominence to future studies. Therefore, the question then is whether Japanese uses pitch range variation to mark discourse salience, in the same way that English (and other languages) uses pitch accent. If so, the prediction is that new information is marked by expanded range in Japanese, while given information is marked by compressed range, regardless of the lexical accentuation of the expressions referring to the discourse entities themselves.

There have been a few studies that have attempted to address this question in Japanese. Sugito examined a limited set of targets situated in short 4–7 sentence written stories [Sugi96]. She found an effect of discourse-*new* (first mention) vs. *given* (later mentions), but only for a subset of the targets in the study. As Sugito points out, her analysis of the given/new distinction is confounded by other factors affecting pitch range in Japanese, such as the effects of downstep, etc.[6] That is, downstepped peaks are by definition lower than non-downstepped peaks, and this phonological effect prevented proper comparisons of the targets. However, in cases where both given/new targets were not downstepped, the height of the new target was somewhat higher than that of the given target (no statistical comparisons are reported and the data are limited). In cases where both were downstepped, there was no difference in pitch range among the two types. This study is a keen reminder of the point discussed in Section 2.3: in analyses of spoken language databases, one cannot make comparisons between tokens, for example comparing the height of peaks A and B, without first knowing what phonological factors gave rise to those peaks. In Japanese, downstepping and the height distinction of H- vs. H* are two phonological factors that must be considered.

A recent series of experiments by Hirose, Sakata and colleagues also examined the effects of discourse information on pitch height in Japanese simulated dialogues (i.e. text read as if in a conversation) [HSOF94, SH95, HSK96].[7] Hirose et al. found that

[6] In Sugito's data, some targets were downstepped due to an immediately preceding accented modifier in the same phrase.

[7] This work uses the Fujisaki model of Japanese intonation, which lacks the notion of *pitch range* per se [FS71, FH84]. Instead, in this model the relevant notion is the 'amplitude' (or height) of the accent commands, which ride on top of separate phrase curves and a variable F0 baseline. For convenience, I will use 'height' to refer to the accent command amplitudes, in much the same way as I have been using 'height' to refer to the pitch range in tone-sequence models like ToBI. However, the two are quite distinct: the height of the topline in the J_ToBI implementation corresponds not just to the height of the accent command in the Fujisaki model, but to that height, plus the height of the phrase curve at a given point in time, in addition to the height of the F0 baseline (which itself can vary from utterance to utterance even for a given speaker, unlike the notion of *reference line* in the J_ToBI implementation (see Section 2.3 above)).

'focused words' are realized by a higher accent command amplitude (i.e. height) in simulated dialog than when produced in isolated sentences out of context, while the distribution of amplitudes of 'de-focused words' in dialogue is the same as that in isolation [HSOF94]. However, while Hirose et al. clearly define 'de-focused' words as those which are "already mentioned in a preceding utterance" [HSOF94, p. 168], their definition of 'focus' as "the noun conveying key information in the case of an answer, or ... the interrogative pronoun in the case of a question" [HSOF94, p. 168] is admittedly vague and not directly equatable with *new* information per se. In addition, their study collapsed downstepped/non-downstepped and accented/unaccented tokens, leaving their conclusions open to the type of criticism outlined above. However, in subsequent studies, Hirose et al. were able to separate out these confounding effects, and examine the discourse effects more directly [SH95, HSK96]. These studies yielded mixed results. Sakata and Hirose [SH95] report increased height of accent command amplitudes (relative to readings in isolation) for downstepped accented words only, for those words representing either (a) "the main target information ... and is not a repetition", (b) "essential for the understanding of the sentence and is a repetition", (c) "essential for the understanding of the sentence but is not a repetition", or (d) when the word "stresses, emphasizes or modifies the understanding of an essential or a target word" [SH95, p. 1008].[8] They did not observe substantial discourse effects for their non-downstepped accented targets, or for their unaccented targets. In addition, while they do distinguish *given* from *new* in terms of a 6 speaker-turn cache, the notions of 'main information', 'essential for understanding', etc., beg for further clarification. A subsequent study by Hirose et al. presents further analysis of their corpus, using a multiple regression analysis to tease apart the effects of discourse, phonological, and part-of-speech factors [HSK96]. This study describes the discourse features in terms of given/new (where *new* is re-defined as that which was not mentioned in the previous 4 speaker-turns) and discourse 'importance' ("the information is necessary to understand the content of the sentence and to make a reply" [HSK96, p. 379]). The analysis showed effects of both given/new and 'importance', though the given/new effect was not robust.

In sum, the distinction between *given* and *new* discourse information can be cued by intonational means, especially in the case of English: new information tends to be pitch accented, while given information tends to be unaccented. As for Japanese, the conclusions are less clear-cut. Japanese uses pitch range as one means to cue intonational prominence, and there is some evidence that new information tends to be marked with expanded range, while given information has a lower range. However, studies have not shown a robust effect of this given/new distinction for Japanese, as in English. One reason for this may be that, in these studies, the notion of *given* in

[8] In this analysis a 'repetition' is defined as a word which had been mentioned within less than 6 previous speaker-turns. Also note that 'accent commands' are used to model both accented and unaccented words in this intonation model.

contrast to *new* has not been systematically defined using a model of discourse organization. In the following sections I will discuss some attempts to characterize discourse *givenness* in terms of global salience, and will relate these new characterizations to the data in English, Japanese, and other languages.

4.2.2 'Given' as 'currently salient' in the discourse

As discussed above, Sugito and Hirose et al. both define *given* (versus *new*) in terms of a text buffer: for Sugito, *given* means those entities which are not the first mention in the story [Sugi96], and for Hirose et al., it means those entities which are mentioned within the previous utterance [HSOF94], or with a 4 or 6 speaker-turn cache [SH95, HSK96]. In contrast, Brown [Brown83] defines given/new in terms of an independent taxonomy of givenness developed by Prince [Prince81]. Both of these approaches hint at a broader definition of *given*: entities which are currently salient in the discourse, by way of their being previously evoked or generally known. However, what is needed is a clearer characterization of what it means to be 'currently salient'. For example, Brown points out an instance in her data in which a 'given' entity is re-introduced into the discourse after some digressions, and is marked by intonational prominence. Using only Prince's taxonomy of givenness, along with a direct mapping of these categories to intonational prominence markings, Brown cannot account for such accentuation of re-introduced entities. However, approaches using the notion of a text cache/buffer of a fixed number of utterances or turns may be able to capture such phenomena. That is, the entity is no longer salient if the number of utterances defined by the cache size have intervened. But what cache size is appropriate? And is the same size appropriate for all discourse situations?

The use of topic-based discourse segmentation is one way to better define what it means for an entity to be 'currently salient', without having to resort to an arbitrary and fixed cache size. For example, Terken examined accent distribution in Dutch housebuilding monologues, using a topic unit defined as a stretch of speech in which a specific piece of the house (the 'topic') is being described [Terk84] (see also [SG94] mentioned above). The discourse entities described within a particular topic unit are taken to be those which are most salient to the discourse participants at that point in time, with the 'topic' being the most salient entity. Terken found that both topics and non-topics are newly introduced using accented full NPs (97% and 81%, respectively). This is consistent with the accentuation of *new* entities in English and other languages. However, Terken observed that the realization of later mentions (within the topic unit) depends on the topic status of the entity: topics are mainly realized by unaccented pronouns (51%), but accented and unaccented full NPs are also found (33% and 5%, respectively). Non-topics, on the other hand, are primarily realized by accented full forms (74%), though unaccented full forms exist as well (18%). These results suggest that there is a general relationship between given/new, as defined by topic unit segmentation, and the intonational prominence marked on those entities via pitch accents. However, this relation is not straightforward, but is obscured by yet another factor: the salience of the topic which is currently being talked 'about'.[9] I will discuss in detail the effect of this *local* focus of attention in Section 4.2.4 below.

Venditti and Swerts studied the effect of given/new defined by topic-based segmentation in Japanese housebuilding monologues [VS96]. They found a tendency for the pitch range to be higher in intonation phrases containing entities which are first mentions in the topic unit, as opposed to those that are later mentions within the unit. This observation holds for both topics and non-topics alike.[10] These results suggest that for Japanese, as well as for English and other languages, a topic-based segmentation of discourse may provide appropriate domains within which to characterize the notion of 'givenness'. Entities defined with relation to such units show systematic patterning in their intonational realizations, albeit via different means (pitch accent in English vs. pitch range in Japanese). In the following sections, I will discuss some classes of 'exceptions' observed in studies using the given/new dichotomy, and show how these classes have been accounted for by studies employing Grosz et al.'s intention-based discourse segmentation and the *global* and *local* attentional state modeling outlined in Chapter 3.

4.2.3 Intonation and global attentional salience

Hirschberg and Pierrehumbert [HP86] suggest that the notions of *given* and *new*, and their relation to intonational prominence, can be explained by a model of global attentional salience such as that proposed by Grosz and Sidner [GS86]. Nakatani provides a detailed formulation of how such attentional salience can account for the accentuation and form of referring expressions in an English spontaneous narrative [Naka93, Naka97a, Naka97b, Naka98]. She observes that entities which are first introduced into the current global focus space (which models the current intention-based discourse segment) tend to be realized with accented referring expressions, while those entities already existing in the space (and hence globally salient) tend to be realized with unaccented expressions. Nakatani also notes that "references to entities that are either in a neighboring focus space on the focus stack, or in the most recently popped focus space, [also] do not require accentual prominence" [Naka97a, p. 149].

[9] Of course, there are also other factors not directly related to discourse which can influence pitch accent distribution. These include part-of-speech, complex NP accentuation rules, persistence of grammatical function and surface position, etc. (e.g. [Hirsch93, TH94], inter alia). This may explain the high percentage of accents on later mentions of topics and non-topics (33% and 74%, respectively) in Terken's study.

[10] Venditti and Swerts do not report statistical comparisons. The means of each group show the *new* > *given* patterning, though the standard errors overlap considerably.

To the extent that intention-based discourse segmentation may in many cases correspond closely to Terken's topic-based segmentation (for example, the topic unit which describes the front door could correspond to a DS with the purpose (DSP) 'instruct how to build the front door'), Terken's results can be directly interpreted in terms of this new approach. In addition, Nakatani's observations using Grosz and Sidner's model can account for two of Terken's 'exceptions': reference to the entity *house* using a non-prominent expression, and decaccenting of some referents when the antecedent is in the previous topic unit. In the first case, Terken notes that "expressions referring to the house itself are often deaccented, even though the house has not been mentioned over long stretches of discourse" [Terk84, p. 280]. One possible explanation for this is that the entity *house* could have been mentioned early on in Terken's discourses, and in a model of these discourses using a focus stack, this entity may continue to reside in a embedding non-immediate focus space which has been pushed down on the focus stack. If this is the case in Terken's data, then Grosz and Sidner's model of global salience would characterize this entity as being in non-immediate global focus, and the global salience would license the use of an unaccented expression to refer to the *house* in subsequent embedded segments. In the second case of 'exceptions', Terken observes deaccentuation of some referents across topic unit boundaries. Though the full details of the data in question are unknown, it is possible that Nakatani's observation that entities in just-popped sister segments are still salient could explain the phenomenon Terken describes. Davis and Hirschberg also note that entities in immediately preceding sister segments are considered salient [DH88].

In addition to the exceptions that Terken notes, a focus stack-like model of discourse salience can account for exceptions such as that noted by Brown (mentioned above), in which a 'given' entity is accented upon its re-introduction to the discourse after a stretch of digressions [Brown83, p. 76]. In Grosz and Sidner's focus stack model, the entity would not be in the current or even immediately preceding focus space, and so is not currently salient when it is re-introduced. Therefore, a pitch accent is used, much the same as if it were a new entity.

Japanese exhibits a similar effect of global attentional salience on the realization of intonational prominence in discourse. Venditti and Swerts reanalyzed their house-building data in terms of Grosz and Sidner's focus stack model, using the topic-based instruction units as a working definition of discourse segments [VS96].[11] They found that entities which are either newly introduced or re-introduced into the focus space are marked by a higher pitch range relative to other entities in the same utterance. In contrast, entities which are already in the immediate (current) focus space, non-immediate (mother) focus space, or the just-popped sister focus space have a lower relative pitch range. These results mirror Nakatani's findings for pitch accent in English [Naka97a].

[11] This reanalysis was reported only in the poster presentation of their paper.

In sum, discourse salience influences the choice of pitch range on referring expressions in Japanese, and of pitch accentuation in English and other languages. As Terken notes, the "speaker's decision to accent or deaccent words depends on whether or not he judges their interpretation to be already available to the listener" [Terk84, p. 271]. Terken chooses to define 'availability' in terms of topic units, while others such as Nakatani have proposed a unified account of accent distribution in terms of the global attentional focus modeled by Grosz and Sidner's focus stack. I will now turn to a description of still another factor which effects pitch accent placement: the *local* attentional salience of a referring expression.

4.2.4 Intonation and local attentional salience

When an entity is first introduced into the global focus space, it is typically realized by an indefinite referring expression, such as *a mango*, and receives prominent intonational marking due to its *newness* in the discourse.[12] Once it is in the space, it is globally salient, and can be referred to by using a definite form (e.g. *the mango*) and non-prominent intonation. In some cases, the entity is also the most salient entity in the local context (the Cb), and can be referred to using a reduced lexical form, such as a pronominal (e.g. *it*). Since the pronoun Center is by definition both locally and globally salient in the current discourse, it is by default realized using non-prominent intonation. However, there is a class of systematic exceptions to this generalization, in which pronouns are realized using prominent intonation. Consider the following excerpt from Nakatani's spontaneous narrative (the underlined pronouns are unaccented, and those in boldface capitals are accented).

> *So Masson became the new curator.*
> *He flies to London and, you know,*
> *he's already met Anna Freud and therefore*
> *he has access to the secret cupboard of Freudian letters*
> *and naturally Anna assumed that uh*
>
>> **SHE [H*]** *was a brilliant woman too —*
>> *she did more a lot of work in child psy- psychiatry*
>> *and psychoanalysis*
>
> *assumed that* **HE [H*]** *would keep this information*
> *you know within the confines of the psychoanalytical group*
>
> *Well, as Masson was studying these letters he realized ...*
> [Naka97a, p. 148]

[12] But see [Brown83] for examples of newly introduced *inferred* entities which are often realized with prominent intonation and a definite expression (e.g. *the driver*).

A number of researchers have observed that accented pronouns are not just unprincipled exceptions, but they serve a specific discourse function: to shift the center of attention to another entity in the discourse model (e.g. [Terk84, HP86, Cahn, Cahn95, Naka93, Naka97a, Naka97b, Naka98], etc.). For example, in his Dutch housebuilding monologues, Terken notes that 'topic shift' may account for the class of 'exceptions' in which accented pronouns are observed. Terken notes that they typically occur at the start of a new topic unit [Terk84, p. 282–283]. Nakatani formalizes such observations of topic shift in terms of the discourse Centering processes outlined in Section 3.4.1 above [Naka97a, Naka97b, Naka98]. She proposes that an accent on a pronoun cues an attention shift to a new Cb. This is illustrated by the two cases in the excerpt presented above. In the first case, the speaker decides to elaborate on an entity (*Anna*) which is not the Cb of the current utterance (Cb=*Masson* at this point), so the speaker starts a sub-segment (push onto the focus stack) which is 'about' this new Cb (*Anna*). The accentuation of the subject pronoun **SHE** indicates this shift in local attention. In the second case, the embedded segment (about *Anna*) is finished, and that focus space pops from the stack, returning to the previous (embedding) DS. The accentuation of the subject pronoun **HE** indicates a shift in attention back to the previous Cb (*Masson*) at the point where the discourse was suspended.

Based on observations such as this, Nakatani formalizes the relation of pitch accentuation in English to both global and local discourse salience, summarized in Table 4.1 (adapted from [Naka97a, p. 143]).[13]

ACC	**full form**	introduce new referent into global focus
UNACC	**full form**	maintain referent in global focus
ACC	**pronoun**	shift local attention (Cb) to new referent
UNACC	**pronoun**	maintain referent in local focus (Cb)

Table 4.1: Nakatani's characterization of discourse salience effects on pitch accent placement in English.

Many studies investigating the intonation-discourse mapping, such as those described above, have focused either on cues to discourse structure, or on effects of global salience like *given* vs. *new*. Nakatani has shown that global focus is only part

[13] For simplicity, Nakatani's distinction between subject and object grammatical role is not included here. The main difference is that subjects serve as preferred Centers (Cps), while objects do not.

of the story. In addition, the speaker's choice to place pitch accents on discourse entities is influenced by the salience of those entities at the local level as well. By using a dynamic model of attentional salience, such as Grosz et al.'s models of global and local focus, Nakatani is able to describe the use of intonational prominence reported in the literature, as well as systematic cases previously considered 'exceptions'. In the next section, I will review some pilot data which shed light on the influence of local focus on choice of pitch range in Japanese.

4.2.5 Intonation and Japanese NP-*wa*

Nakatani and others have observed that non-prominent intonation on a referring expression serves to maintain the referent in global or local discourse focus. In English, unaccented full NPs mark globally salient entities, while unaccented pronouns mark (continuation of) salience at the local level. In Japanese, global discourse salience has been shown to influence the speaker's choice of pitch range on a referring expression. But what about the effects of local discourse salience?

As described in Section 3.4.2, locally salient entities are typically realized as zero pronominals in Japanese, whose interpretations are governed by the principles of Centering Theory. Unlike the case of English pronouns, zeros are not phonologically realized in the surface form of a Japanese utterance, and thus can not provide any data for a study of intonational marking of discourse salience. However, Japanese does provide morphological means, via the postposition *wa*, to mark certain overt NPs as what the utterance is centrally 'about' (see Kuno's 'thematic *wa*' [Kuno73]).[14] Kuno equates the function of topic-marked NP-*wa* to that of zero pronominals [Kuno72], and Walker et al. take this to motivate the *Center instantiation* of NP-*wa* as Cb discourse segment-initially. Therefore, for the purposes of investigating the intonation-discourse mapping, NP-*wa* can be used to study the influence of local salience on intonational realization, specifically pitch range variation, in Japanese.

One prerequisite of (thematic) *wa*-marked entities is that they must be salient to the hearer in some way. That is, they must be entities which are either *unused*, *textually- or situationally-evoked*, or *inferred* as defined by Prince's [Prince81] taxonomy. These are the categories that license use of a definite expression like *the mango* in English. However, these categories do not map exactly onto distinctions in discourse salience modeled by global focusing. For example, while *textually-evoked* entities are likely to be globally salient (if evoked by previous mention within the same focus space) and realized by non-prominent intonation, *unused* or *inferred* entities can be either new or old in the global focus stack, and so may differ with respect to their intonational marking. Indeed, Brown found that in her English data, *inferred* entities are marked by intonational prominence, just as are *brand-new* entities

[14] Throughout this thesis, 'NP-*wa*' refers to Kuno's 'thematic *wa*' unless otherwise specified.

[Brown83]. Therefore, it is not likely that we will be able to predict the intonational characteristics of NP-*wa* expressions directly on the basis of the categorization presented by Prince's taxonomy. Rather, we will have to refer to the entity's status with respect to the global and local focus of attention as described in Grosz et al.'s model.

There have been few studies which have examined the intonational characteristics of NP-*wa* specifically, and in general these studies have not controlled for the many factors, both discourse and otherwise, that are known to affect intonational realization in Japanese. Kuno has observed that "while noun phrases preceding the thematic *wa* do not receive prominent intonation, those preceding the contrasting *wa* do" [Kuno73, p. 47]. Finn also observed this distinction between the two types of NP-*wa* [Finn84]. In addition, she also observed that NP-*wa* exhibits a larger F0 fall (from peak to subsequent valley) than cases of NP-*ga*, suggesting that NP-*wa* is intonationally more prominent than the corresponding NP-*ga*. However, Finn's analysis is confounded by a number of other factors that are known to affect F0 height, including the lexical accentuation of the target NP, prosodic phrasing and downstep, sentence-internal position, vowel height, etc. Therefore, no definite conclusions can be drawn from her results.

Pilot data

In a pilot study to this thesis, I examined the variation in pitch range across phrases in a narrative monologue database. The monologues were elicited by asking the speaker to narrate a story using a sequence of hand-drawn pictures used as prompts. This elicitation method minimizes the memory and cognitive load on the speaker (unlike the housebuilding task, which involves many on-line decisions), resulting in fluent spontaneous discourses containing few hesitations or other disfluencies. Then, after the spontaneous monologue is recorded, it can be transcribed and used again for studies of read speech. Figure 4.1 shows data collected from one of the read monologues.

The figure shows a Classification and Regression (CART) tree [BFOS84, Ril92] which models the variations in pitch range in one speaker's monologue. The tree and features shown here have been truncated to save space. Splits in the tree are determined by which combinations of features and feature values will minimize the prediction error after that split. The hertz value in each square is the average difference between the observed F0 peak value and the peak value that is predicted by a 'default' pitch range model. The default model includes variables such as typical initial values for the pitch range topline and reference line, the amount of reduction at each downstep, and the relation of H- to H* heights (these are mostly speaker-specific values, and were extracted for this speaker from a standard set of read sentences). Using such a model, we are able to avoid many of the confounds existing in Finn's analysis. The data presented in the figure represent deviations from predicted values

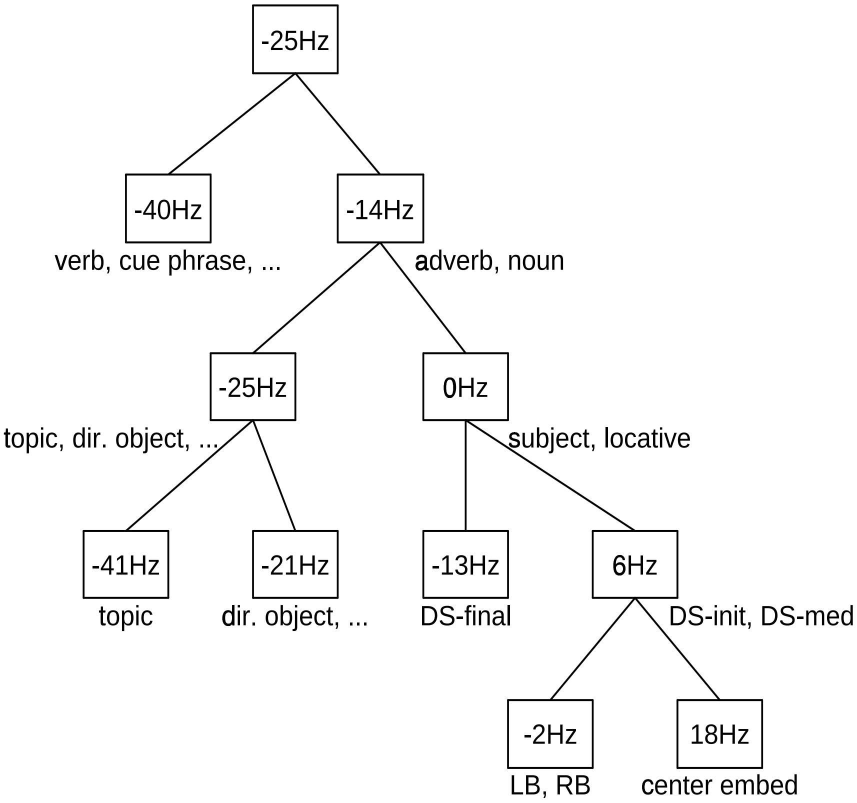

Figure 4.1: CART tree showing a model of pitch range differences (observed minus predicted peak heights) according to tagged features in a read monologue.

due to syntactic and discourse effects, without known influences from the phonetic factors.

There are important deviations from the predicted value, in both directions. Cue phrases (such as *tsugi ni* 'next') and verbs are on average produced in a lower range than predicted (the peaks are 40 Hz lower), while adverbs and nouns pattern differently by being produced in a higher range (albeit still lower than predicted, by 14 Hz). Among nouns, this analysis shows that *wa*-marked topics and objects in this discourse have a lower range, while *ga*-marked subjects and locative NPs are produced right at the predicted height. Among this latter subset of noun phrases, NPs that are final to the discourse segment are lower than DS-initial or DS-medial ones, and so on.

With regard to topic marking, this analysis suggests that NP-*wa* phrases are realized in a very low range: more than 40 Hz below the predicted value. Such a low F0 range on NP-*wa*, in comparison to NP-*ga*, runs counter to Finn's observations. This intonational non-prominence on NP-*wa* is hypothesized to result from at least two influences: the marking of global attentional salience and the marking of local attentional salience of the entity in the discourse. However, since the NP-*wa* phrases in this pilot database are tagged only for topic-hood (via morphological marking), and not for local salience relations, it is difficult to tease apart the contributions of global vs. local salience to the realization of the NP-*wa* targets. That is, although *wa* marks the entities as locally salient, there is no tagging of Center shifts vs. continuations, a distinction which is known to affect intonational realization (see Section 4.2.4).[15] Previous studies have suggested that global salience licenses reference using non-prominent intonation on entities (*wa*-marked or otherwise) in Japanese, though no studies have looked specifically at effects of local salience, or of salience *relations* on NP-*wa*.

However, a closer examination of the F0 deviations *within* the NP-*wa* category in these data gives a first hint that local attentional salience relations may indeed play a role in determining the intonational prominence, separate from global salience. Consider the following excerpts from the monologue, one near the beginning of the discourse which describes one of the characters named Mayumi, and the other from

[15] It has just been brought to my attention that a recent statistical analysis of spontaneous dialogues by Fry [Fry00] supports Finn's conclusions that NP-*wa* phrases are realized as relatively *more* prominent than NP-*ga*. However, there are a numbers of differences between Fry's study and the data reported here. For example, Fry did not explicitly code for 'contrastive' vs. 'thematic' NP-*wa*, a distinction which Kuno has claimed is marked by intonational means [Kuno73] (that is, contrastive NP-*wa* is realized in an expanded range). The analysis in this thesis restricts attention to 'thematic' NP-*wa* only. In addition, since Fry's database is not tagged for discourse structure, global salience, or local salience relations, it is difficult to determine the extent to which these factors influence the intonational realization of the NPs (results reported in Chapters 6 and 7 of this thesis suggest that they do play a significant role). It will be interesting in future studies of spontaneous dialogue to code for such discourse features.

further along in the discourse when Mayumi meets her friend Aya (the *wa*-marked targets are underlined):

> *... mado-no soto-o minagara,*
> *tenisu-ka jogingu-ka nani-ka-o shi-ni ikô-to kangaemashita.*
> *chôdo rôrâburêdo-o hajimeta-bakari datta node,*
> *rôrâburêdo-ni shiyô-to omoitsukimashita.*
> *soko-de Mayumi-wa Maruyamakôen-ni dekakemashita.*
> *Maruyamakôen-de rôrâburêdo-de tondari ...*
> *...*
> *sô shite iru uchi-ni gûzen chikaku-o tomodachi-no Aya-ga*
> *tôrikakaru-no-o mitsukemashita.*
> *ôgoe-de Aya-o yobimashita.*
> *Aya-wa jitsu-wa yojikanmae-ni Maruyamakôen-ni kite imashita.*
> *e-o kaku tame-ni funsui-no mae-de kyanbasu-o hiroge ...*
>
> ... While looking out the window,
> she [Mayumi] thought about going to play tennis or jogging.
> Since she [Mayumi] had just started rollerblading,
> she [Mayumi] came up upon the idea of going rollerblading.
> So Mayumi set off to Maruyama Park.
> In Maruyama Park, she [Mayumi] skated around on the rollerblades ...
> ...
> While doing that, she [Mayumi] suddenly found her friend Aya walking nearby.
> She [Mayumi] called out to Aya.
> Aya had been in Maruyama Park since about four hours ago.
> She [Aya] set up her canvas in front of the fountain
> in order to paint a picture ...

In these two excerpts, there are two NP-*wa* targets. The first one, *Mayumi-wa* represents a *continuation* of the local Center of attention from the previous utterance, and is realized in a low range (deviation from predicted = -30 Hz). In the second case, *Aya-wa* serves to *shift* the local Center of attention from the previous Cb (*Mayumi*), and is realized in a relatively higher range (deviation = -10 Hz).[16] So, while both of these NP-*wa* targets are globally salient in the discourse (they appear in the previous utterance), the difference with respect to the speaker's *local* attention, specifically the local salience relations, may be what causes the 20 Hz difference in pitch range between the two targets. If this is the case, this result would mimic that observed

[16] Note that while *Mayumi* is lexically unaccented and *A 'ya* is accented, the deviation-from-predicted measures already take this into consideration, and thus the differences in accent do not confound the comparison.

for English by Nakatani and others: intonational non-prominence serves to maintain the entity in local focus, while intonational prominence serves to cue a shift in local attention.[17] Of course, these pilot study observations for Japanese require further experimental validation. Therefore, this effect of local attentional salience relations on Japanese NP-*wa* targets, and also the effect of global salience, are the main issues that the remainder of this thesis examines in detail.

4.3 Open research questions

Large spoken language databases are a valuable resource for research on the intonation-discourse interface, provided that the databases are tagged for both intonational and discourse structures. However, in this tagging, it is often difficult to know which intonation or discourse features matter, and this has been the pitfall of many previous studies. In this chapter, I have reviewed several attempts to relate intonation to discourse structures, and have shown that studies which synthesize the theory of intonation described in Chapter 2 with the theory of discourse described in Chapter 3, have yielded promising results in this area. In languages such as English, many studies have shown that variations of pitch range can cue discourse segmentation, by marking the edges of segments. The salience of discourse entities, at both the global and local levels, is cued by the use of pitch accents. In Japanese, on the other hand, pitch accent does not have a discourse function, so pitch range is used for both purposes. Studies have shown that range variation can mark segment edges, and there has been some evidence which suggests that Japanese also uses pitch range to cue salience of discourse entities at the global level. As for the effects of local attentional salience, the jury is still out.

This thesis further investigates the relation of intonation to discourse structures in Japanese. Specifically, it is an attempt to explore the intonation-discourse mapping using the structures provided by the J_ToBI model of intonation and Grosz et al.'s model of discourse structure and attentional salience. The goal of the thesis is not only to confirm findings of previous studies using these new models, but to expand on these findings to investigate aspects of the mapping which have yet to be studied. The remainder of the thesis describes a controlled experimental production study which was designed to address the following open research questions in Japanese:

[17] A study by Nakajima and Tsukada also suggests that increased pitch range is correlated with 'topic shift' in Japanese dialog [NT97]. In this study, 'topic units' were defined as those stretches of dialogue which pertain to a particular communicative goal. They found an increased pitch range at the start of topic-shifting units, relative to the range at the start of topic-continuing units. These results may support the observations reported here for the Japanese pilot data, though Nakajima and Tsukada did not control for global discourse salience or part-of-speech in their analysis, so I am hesitant to make a direct comparison to these data.

- **Is pitch range variation correlated with intention-based discourse segmentation?** Previous studies have shown effects of written 'paragraph' structure [Ven96] and topic structure [VS96] on pitch range variation in Japanese, but what about discourse segmentation based on speaker intentions?

- **Can pitch range variation cue the global attentional salience of a discourse entity?** Previous studies have shown that both distinctions of 'given' vs. 'new' [Sugi96, HSOF94, SH95, HSK96] and global salience as defined by Grosz and Sidner [VS96] can to some extent influence the intonational realization of referring expressions in Japanese, though such analyses have not teased apart these effects from discourse structure effects (above) and possible local salience effects.

- **Is the choice of pitch range on 'discourse-given' referring expressions influenced by the location of the antecedent with respect to the focus stack?** Previous studies have suggested that non-prominent intonation is used when the antecedent is either (a) in the same focus space, (b) in a non-immediate (mother) space on the stack, or (c) in a just-popped sister space [Naka97a, Naka97b, VS96]. This study seeks to confirm these findings for Japanese, while separating out possible confounding effects of discourse structure and local salience (as mentioned above).

- **Does the local attentional salience of the antecedent affect the intonational realization of a discourse entity?** Previous studies have found that speakers are more likely to use non-prominent intonation in subsequent mentions of entities which are 'topics', in comparison to subsequent mentions of 'non-topics' [Terk84]. Is this effect due to local salience of topic entities themselves, or might it also be due to the salience characteristics of the antecedent? This study examines the intonational realization of referring expressions which have Centered vs. non-Centered antecedents, in order to examine any such potential effects.

- **What effect does local salience have on the choice of pitch range?** Previous studies have not examined the effect of local discourse salience to the exclusion of other confounding factors. This study compares topic-marked NP-*wa* and direct object NP-*o* discourse entities.

- **Does Center maintenance or shift influence choice of pitch range?** Previous studies have shown that pitch accentuation in English and other languages can cue the speaker's intent to maintain or shift the local Center of attention (e.g. [Terk84, Cahn, Cahn95, Naka97a]). A pilot study to this thesis suggests that Japanese may use variations in pitch range in much the same way (see Section 4.2.5).

In the next chapter, I will describe the design, recording, and analysis of the database constructed to address these open research questions. In Chapter 6 I will review results from the analysis of discourse structure and global salience effects on intonational realization of targets, and in Chapter 7 I will review the local salience effects. I will conclude with a unifying discussion of the experimental findings in Chapter 8.

CHAPTER 5
DATABASE DESIGN AND ANALYSIS

This chapter describes in detail the discourse structure and placement of target phrases in a Japanese read speech database, which was designed specifically to address the open research questions outlined in Section 4.3. In studying the intonation-discourse mapping in Japanese discourse, the ultimate goal is to be able to describe the relation between the two linguistic structures based on large spoken language databases containing either read (e.g. news stories) or spontaneous (e.g. conversation) naturally-occurring discourse. However, such databases are not ideal for our purposes here. That is, we are interested in testing specific hypotheses about how intonation relates to discourse structures, and thus we need to be certain that the intonation and discourse configurations that are relevant for such hypotheses occur with adequate frequency in the database. Although naturally-occurring data contains a wide range of variation in both the intonation and discourse domains, without a sufficiently large amount of tagged data, we cannot guarantee that the configurations of interest will be well-represented.[1] For this reason, I chose to examine a database of semi-controlled read speech, which I constructed with the generous help of Yuki Hirose, a native Japanese speaker.[2] I use the term 'semi-controlled' to highlight that only certain aspects of the discourses were controlled for the experimental analysis. As I will describe in detail in Sections 5.1 and 5.2 below, only the intention-based structuring and the locations of selected target phrases were controlled, and the content and form of the remaining discourse utterances were free for the native speaker to improvise. This allowed for the controlled discourse (and intonational) configurations of interest to be situated in an otherwise naturally written discourse.

Nine different discourses were constructed which describe steps for the preparation of various Japanese and Western food dishes: for example, a Hawaiian-style breakfast (see Figure 3.1), miso soup, grilled fish, etc. This type of data is comparable to the task-oriented instruction discourses described by Grosz and colleagues (e.g. [Grosz77, GS86, NGAH95]). All discourses are identical in their overall discourse structure, following Grosz and Sidner's intention-based method of discourse segmentation. In Section 5.1.1 I will discuss how the intention-based segments were defined in such

[1] For example, see [vS94] for a discussion of the problem of data sparsity in acoustic modeling.

[2] Yuki was a graduate student in the Department of Linguistics at CUNY at the time, specializing in syntax and psycholinguistics. I owe a debt of gratitude to her for many hours of help.

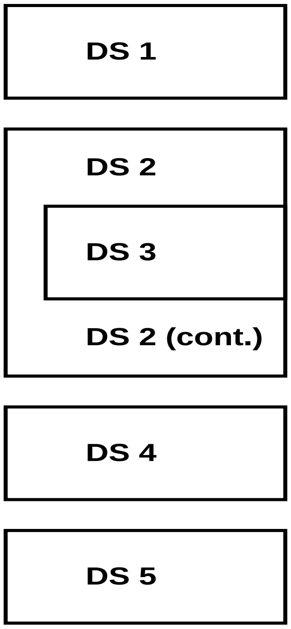

Figure 5.1: Schematic representation of the structure of each discourse.

constructed data. The discourses differ with respect to the distribution of target phrases in each: targets are situated in key locations throughout the discourses, as described in Section 5.2 below. Japanese transcripts of the nine discourses are given in Appendix A.

5.1 Structure of discourses

Figure 5.1 shows the intention-based structuring of the discourses. Each discourse has an overall purpose, and four main sub-purposes corresponding to the purposes of DS1, DS2, DS4 and DS5. These segments are in a sister relation to each other. In addition, DS3 is an embedded segment whose purpose contributes to the purpose of the embedding DS2. In terms of the global focus stack, the space corresponding to DS2 is pushed down on the stack when DS3 is opened, and is then 'resumed' when DS3 is popped from the stack.

5.1.1 Intention-based segmentation in constructed discourses

Discourse segmentation using Grosz and Sidner's [GS86] intention-based approach is usually performed by a third party labeler on naturally-occurring written, read or spontaneous discourse (e.g. [GS86, GH92, PL93, HN96, Naka97b], etc.). In this case, the segmentation is based on what the labeler estimates the speaker/writer has intended by producing the discourse in a given manner. In the case of a constructed written discourse, in contrast, the intentions of the discourse-producer are necessarily known. However, how do we know this structure will be perceived by a reader (i.e. by the native speaker who produces the read speech) as it was intended? Again, this is the job for an independent labeler, just as in the case of naturally-occurring discourse. In the construction of the discourses for this database, I employed three native labelers to confirm the discourse structuring that was intended.

The first step in constructing the discourses was to design a template of intentions for each discourse, similar to the `WHY?` outline of purposes shown in [NGAH95, p. 5, Figure 5]. The location of key target phrases was also specified in relation to this structure of purposes. Then, I employed the help of a native speaker (YH) to fill-in the rest of the text, to produce a fluent description of the recipe. The exact content and form of the text was revised a number of times by a collaborative effort between YH and the author (who is also a speaker of Japanese). The purpose of these revisions was to make the discourse segments and purposes as clear to the reader as possible.

After a draft of the discourses was complete, two native labelers (TK and MF) segmented the 9 discourses following the annotation guidelines developed by Nakatani et al. [NGAH95]. Neither labeler was aware of the discourse design, target placement, or hypotheses of the experiment. The text that was given to these labelers was formatted with one grammatical sentence per line, with no separations between lines. As expected, there was not total agreement among the two labelers, nor between the labelers' segmentations and the intended segmentation (though segmentations of the fruit salad discourse were remarkably similar). There were notable differences in the granularity of the segmentations: TK tended to mark many sub-purposes with a high degree of embedding, while MF preferred larger segments with a flatter structure. These judgments were considered in a further revision of the discourses. Specifically, additional cue phrases and some rewording of the non-target phrases were used to clarify the intended intentional structure.

One possible way of constructing discourses which are perceived to have the exact intentional structure as intended by the database design is to conduct an extended iterative process of revision, independent labeling, more revision, etc. However, this is not only time-consuming and costly, but there is no guarantee that if a new labeler were added, she wouldn't have a slightly different judgment of the writer's intentions. That is, it is not likely that we could hone in on a single unanimous segmentation using this method. Therefore, instead of many revise→label→revise iterations, I decided to 'induce' the perceived segmentation in a very simple way — by presenting the revised written discourses to native speakers using a formatted text with whitespace indicating 'paragraph' breaks. It is important to note that these 'paragraphs' are based on intention-based segmentation, and they are not necessarily based on any other rules for paragraphing that may be used in written Japanese (though the two definitions of 'paragraphs' may often coincide). The embedded segment (DS3) was cued in the formatted text by an indented paragraph block which had a illustrative flag next to it reading "cooking tip". Such cooking tips are considered to be embedded segments in that they provide "extended details on an object or subpart of a process relevant to the supported WHY?, or [they] could provide general advise on some aspect of the supported WHY?" [NGAH95, p. 8].

A final issue to consider when examining intonational cues to structures in such constructed discourses is the extent to which readers may perceive segment boundaries *within* the main segments which are represented by paragraph blocks. That is, clearly readers will perceive a segment boundary between blocks (this is the reason for such formatting), but in addition, readers might perceive discourse breaks within the blocks as well. This may affect the intonational realization of targets at such locations. For example, a target which is medial in the intention-based paragraph block may be perceived as initial to a sub-segment within the block, thus possibly affecting its intonational realization (see Section 4.1 for a discussion of DS-initial effects on intonation). In order to be aware of potential effects of this type, the final formatted discourses were subject to yet another pass of labeling by a separate native speaker judge (MU), this time after all the recordings had been completed. This labeler was first trained in discourse segmentation methods using the guidelines in [NGAH95], and then was given pre-segmented text (the formatted text represented in terms of WHY? labeling, just as in Appendix A) and was asked to indicate locations of additional potential DS-boundaries with two degrees of certainty: either she judged that there *may* be a perceived break there, or there *surely* should be a break there. As outlined in Section 3.2.1, these are cases in which labelers may judge a sub-segment within a larger DS. So, although this last set of judgments was done post hoc and did not influence the construction of the discourses themselves, it does provide flags of a few cases in which targets may have been interpreted by readers differently from how they were intended to be interpreted. I will return to mention these cases in the presentation of the results in Chapters 6 and 7.

5.2 Target positions

Target noun phrases were placed in strategic locations in the nine discourses, as described in the sections below.[3] The targets represent entities which are ingredients (wine, mango, etc.) or implements (cooking pot, cutting board, etc.) used in the recipes. In order to ensure proper intonational comparison, targets were controlled for their prosodic characteristics and contexts. All targets are nouns 2–4 morae in length, followed by the postposition *o* (object) or *wa* (topic), and are lexically accented on the vowel /a/ in their 1st or 2nd mora: for example, *a'wabi-o* 'abalone-OBJ' or *na'su-wa*

[3] Table A.1 in Appendix A summarizes the identity and exact location of each target in the discourses. Discussion in this chapter focuses on properties of the discourse context in which the targets were placed.

‘eggplant-TOP’, etc. Targets were placed in intonation phrase-initial position, which are all also sentence-initial.[4] Carefully controlling the intonational properties of the targets like this allows for direct comparison of targets in different discourse positions, without possible confounds from phonological factors such as downstep, lexical accentuation, etc. (see discussion of potential confounds in the analyses presented in Section 4.2). In addition, this control also allows for comparisons using different lexical items in the various discourse locations. Doing so enables us to scatter many targets within the few separate discourses, without worry that multiple repetitions of any one entity will confound the analysis of ‘given’ vs. ‘new’ or global discourse salience.

5.2.1 Discourse structure targets

In order to examine the effect of intention-based discourse structure on pitch range variation in Japanese, targets were situated in discourse segment-initial, medial and final positions. One particular segment, DS4, was chosen to be the representative DS to examine the DS-internal position contrast, since this DS is neither absolute discourse-initial nor discourse-final, nor is it adjacent to an embedded segment (which might possibly have an effect). The discourse structure targets are of the form NP-*o* (in fact, all targets in this condition are *mana'ita-o* ‘cutting board-OBJ’), placed in sentence-initial position of DS-initial, medial and final sentences (the DS-final sentence was equivalent to the DS-final intonation phrase). In other words, in terms of their intonational context, these targets are initial to intonation phrases which are either the initial, medial, or final phrase in DS4. In addition, all targets in this condition are neither globally salient (they are the first mention in the discourse), nor locally salient (they are not the Cb of the utterance). In addition to these targets in DS4, *mana'ita-o* targets were also placed in initial position of the other discourse segments: DS1, DS2, DS3, DS5, and also at the start of the resumed DS2 (which is technically DS-medial position). Table 5.1 summarizes the properties of the discourse structure targets and the additional *mana'ita-o* targets.[5]

5.2.2 Global attentional focus targets

In order to examine the effects of global attentional salience on pitch range variation in Japanese, additional targets were placed at strategic locations in the discourses. Like the discourse structure targets, these targets are also of the form NP-*o*, located

[4] Most targets were also uttered with an intonation phrase break following the NP, though there were some exceptions to this.

[5] Table A.1 in Appendix A provides information about the exact location of all targets: the discourse, discourse segment, and sentence in which each occurs in the database design.

target	surface form	location of antecedent	Cb of utterance?	DS-internal position
`init`	NP-*o*	(none)	no	initial (DS4)
`med`	NP-*o*	(none)	no	medial (DS4)
`fin`	NP-*o*	(none)	no	final (DS4)
`initDS1`	NP-*o*	(none)	no	initial (DS1)
`initDS2`	NP-*o*	(none)	no	initial (DS2)
`initDS3`	NP-*o*	(none)	no	initial (DS3)
`initDS2res`	NP-*o*	(none)	no	medial (DS2res)
`initDS5`	NP-*o*	(none)	no	initial(DS5)

Table 5.1: Summary of discourse structure target positions.

in sentence-initial IP-initial positions, and do not function as the Cb of the utterance. However, in contrast to the discourse structure targets, the global focus targets were placed in DS-medial utterances only.

Targets in this set differ with respect to the location of their antecedent in the previous discourse context. That is, with the exception of the `new` target, all global targets can be interpreted as definite referring expressions (e.g. *the mango*) whose antecedent has been introduced at some point prior in the discourse. The configurations investigated in this study are summarized in Table 5.2.

target	surface form	location of antecedent	antecedent→target DS locations	Cb of utterance?	DS-internal position
`new`	NP-*o*	(none)	(NA)	no	medial
`non-adj`	NP-*o*	non-adj. popped sister DS	DS1→DS4	no	medial
`adj`	NP-*o*	immed. adj. popped sister DS	DS1→DS2	no	medial
`sameRES`	NP-*o*	same DS (before embed.)	DS2→DS2res	no	medial
`same`	NP-*o*	same DS	DS4→DS4	no	medial

Table 5.2: Summary of global focus target positions.

In the case of the `new` target, there is no prior mention of that entity in the entire discourse. All of the other global targets are considered discourse-‘given’, in that they are not the first mention in the discourse. In both the `non-adj` and `adj` targets, the antecedent is mentioned in a previous sister DS. The difference between these two target types is that in the `adj` case, the focus space containing the antecedent is popped from the stack immediately preceding the current space (which contains the target), while in the `non-adj` case, the space containing the antecedent had been popped from the stack previously. The `same` target has as its antecedent an entity mentioned in the same discourse segment (though not in the immediately preceding utterance). Likewise, the antecedent of the `sameRES` target is also mentioned in the same DS, but at a point prior to the push of the embedded segment. That is, the `sameRES` target is in the resumed portion of DS2, while its antecedent is in the portion before the embedded DS3. These configurations of antecedent-target pairs were designed to investigate the notion of *global salience* in detail, and to examine how differences in salience can potentially affect the choice of pitch range on a referring expression.

In addition to these global focus configurations, three additional configurations were considered. In order to examine any potential effects of the discourse salience of the antecedent itself on the realization of the target phrase, the experiment also varied the status of the antecedent with respect to Centering. In the cases outlined above (summarized in Table 5.2), neither the targets nor their antecedents function as the Cb of the utterance in which they occur. However, an additional set of targets was also considered, in which the antecedent of the `non-adj`, `adj` and `sameRES` targets functions as the Cb of its utterance (a Cb which is in fact continued for two or more utterances). These new targets will be referred to as `non-adj(C)`, `adj(C)` and `sameRES(C)`: the (C) highlights the fact that their antecedents function as the Center of the successive utterances in which they occur. However, as with the other global focus targets, the target itself is a non-Centered entity.

5.2.3 Local attentional focus targets

In order to examine the effect of local discourse salience and salience relations on pitch range variation in Japanese, this study also included a set of targets of the form NP-*wa*, a topic-marked noun phrase. Each of these local focus targets was placed in sentence-initial, IP-initial position, and in most cases the target is also DS-initial, as will be described below. The targets are distinguished with respect to two variables: the Centering transition between the previous and target utterance, and the relation of the target DS to the immediately preceding DS. These variables will be discussed in detail below. But first, it is important to reiterate the role that NP-*wa* plays as the utterance Cb, and also to address the controversial question of Centering across discourse segment boundaries.

NP-*wa* as the discourse Center

Kuno claims that using a zero pronoun in Japanese is equivalent to marking an entity with the postposition *wa*, in that both represent the 'theme' of an utterance, or what it is centrally ‘about’ [Kuno72]. Walker et al. take this observation as motivation for their *Center instantiation* process, by which NP-*wa* is automatically considered the Center (Cb) of discourse segment-initial utterances [WIC94]. In the current study, most of the NP-*wa* target phrases are situated in DS-initial position, and thus are considered the Cb (consistent with either Kuno's or Walker et al.'s account).[6] That is, each NP-*wa* target is taken to represent the local center of attention at that point in the discourse.

Centering across segment boundaries?

Centering Theory was originally intended to model local coherence *within* discourse segments [GJW95]. That is, the Centering structures, transitions and principles, summarized in Section 3.4.1, assume sequences of utterances in which no DS boundary intervenes. One reason for restricting the domain to the discourse segment is because Centering is meant to describe the attentional salience of discourse entities at a more local level than operations on the global focus stack can describe. While the focus stack models the salience of entities at the global level (and the use/interpretation of (in)definite full NPs), Centering models salience at the local level (and the use/interpretation of pronominal forms).

However, this segment-internal assumption has been relaxed in recent studies (e.g. [Walk98, Pass98, GS98, DiEug98]). As more naturally-occurring data with more complex hierarchical structures are collected and analyzed, there is an increasing recognition of the need to describe local coherence *across* DS boundaries as well as within the segments. Walker cites several reasons for abandoning the DS-internal restriction [Walk98]. For instance, she gives a number of naturally-occurring examples of cases in which discourse Centers are continued across DS boundaries by the use of full NPs as well as by pronouns. She claims that such transitions can occur regardless of the hierarchical discourse configuration, such as in sister segments, DS embedding, or DS pops (resumption of previous DS), etc. In addition, studies of intonational correlates of global salience also suggest that entities in just-popped segments can retain their (global) salience (see [DH88, Naka97a, VS96], discussed in Section 4.2.3). If such referents in adjacent segments can remain globally salient, then they could possibly serve as antecedents for pronominal reference, as observed in Walker's data. In a recent paper, Grosz and Sidner note that the interaction of global and local attention is still an open research question: they suggest that forward- and backward-looking functions implicit in the structures of Centering (Cfs and Cb) could possibly be carried over certain DS boundaries [GS98]. So, although Centering processes have

[6] As I will discuss in detail below, two of the NP-*wa* targets are located in DS-medial position. In one case (`contSAME`), the NP-*wa* continues the Center from the previous utterance, and so is considered Cb for that reason. In the second case (`contRES`), the NP-*wa* continues the Center of the suspended DS2 after the pop of DS3, and is in ‘initial’ position in the resumed DS2.

traditionally been thought to hold only within discourse segments, recent research has suggested that this may not be necessarily so. For the purposes of the current experiment, I assume that Centering structures and transitions do indeed apply across DS boundaries, and I will show that the intonational variation of NP-*wa* targets in fact *requires* such an assumption.

Centering transitions and hierarchical relationships

This experiment examines the pitch range of NP-*wa* in a variety of discourse configurations. Two factors are varied: the Centering transition between the previous and target utterance, and the relation of the target DS to the immediately preceding DS. As outlined in Section 3.4.1, the transitions currently defined by Centering Theory are *continue*, *retain*, *smooth shift*, *rough shift*, and *null* (for discourse-initial utterances). Since the local focus (Cb) targets in this experiment are all NP-*wa* topics, and as such represent the highest-ranked entity on the Cf list of U_n (the Cp), *retain* and *rough shift* transitions (where $Cb_n \neq Cp_n$) by definition do not apply. Therefore, *continue*, *(smooth) shift*, and *null* transitions are examined. A *null* Center transition occurs when there is no previous existing Cb (i.e. in absolute discourse-initial position). Center *continuation* occurs when the Cb of the previous utterance (U_{n-1}) is the same as the Cb (NP-*wa*) of the current utterance (U_n). Center *shifting* (in general) happens when the Cb(U_{n-1}) is not the same as Cb(U_n). There are two cases of center shifts in the database: *smooth* vs. *'hard'*. A *smooth shift* transition occurs in the database when an entity is introduced as a direct object NP-*o* in U_{n-1}, and is subsequently referred to as the Cb target NP-*wa* in U_n. This transition is defined just as it is in the Centering literature. However, *'hard' shift* is not defined in the literature. This transition is technically a subtype of *smooth shift*, because $Cb_n \neq Cb_{n-1}$ and $Cb_n = Cp_n$. However, in this case, the Cb (NP-*wa*) has not been mentioned previously in the discourse, and as such is a newly instantiated Cb. This type of transition is distinct from a *null* transition, because the Cb has indeed been shifted from a previous entity, and it is also different from Di Eugenio's [DiEug98] *center-establishment*, since the target Center was not previously in global focus.[7] Therefore, I have chosen to call this transition *'hard' shift*, not only to distinguish it from the well-known *smooth shift*, but also to emphasize that it is still considered to be a Center *shift*.

In addition to variation of Centering transition, each target NP-*wa* is also characterized by the relation that the purpose of its DS (the 'target DS') has with the purpose of the previous DS. For example, the target segment may be a *sister* DS, or may be *embedded* with respect to the previous DS, etc. Table 5.3 summarizes

[7] Note that Di Eugenio's *center-establishment* [DiEug98] is different from Kameyama's *center-establishment* [Kame85, Kame86, Kame88], and is also different from Walker et al.'s *center-instantiation* [WIC94].

the local focus targets examined in this study, and their characterization in terms of Centering transitions and hierarchical structure.

target	surface form	location of antecedent	Cb of utterance?	transition type	relation to prev. DS	DS-internal position
nullINIT	NP-*wa*	(none)	yes	null	disc-init	initial (DS1)
hardSIS	NP-*wa*	(none)	yes	hard shift	sister	initial (DS2)
smoothSIS	NP-*wa*	immed. prev. DS	yes	smooth shift	sister	initial (DS2)
smoothEMB	NP-*wa*	immed. prev. DS	yes	smooth shift	embedded	initial (DS3)
contSIS	NP-*wa*	immed. prev. DS	yes	continue	sister	initial (DS2)
contEMB	NP-*wa*	immed. prev. DS	yes	continue	embedded	initial (DS3)
contRES	NP-*wa*	same DS (before embed.)	yes	continue	resumed	medial (DS2res)
contSAME	NP-*wa*	same DS	yes	continue	(NA)	medial (DS1/2)

Table 5.3: Summary of local focus target positions.

Figure 5.2 gives schematic representations of these target configurations. Here, 'X' represents the target entity, and 'Y' some other discourse entity. The target is NP-*wa* ('X-wa') in all cases. The nullINIT target is an entity which is introduced discourse-initially. It has no antecedent, and there is no Centering transition defined in this case. The contSIS, smoothSIS and hardSIS targets are all positioned initial in a DS which is in a sister relation to the previous DS. The difference between these targets is the Centering transition: in contSIS the Cb is continued to the target utterance, in smoothSIS the Cb is shifted to the entity which was the direct object ('X-o') of the linearly-recent utterance (in the immediately preceding DS), and in hardSIS the Cb is shifted to an entity not previously mentioned in the discourse. The contEMB and smoothEMB targets have similar transitions as contSIS and smoothSIS, but in these cases the target DS is an embedded segment.[8] The contRES target continues (or 'resumes') the Cb from the last utterance (hierarchically-recent but not linearly-recent) of the suspended DS2 after the pop of the focus space corresponding to DS3. And finally, contSAME continues the Cb from the previous utterance within the same DS.[9] By placing NP-*wa* targets in strategic discourse positions such as this, it will be possible to examine the effect that the local discourse context has on intonational variation in Japanese.

[8] A **hardEMB** target was not included because of the difficulty in constructing a discourse situation in which an embedded DS starts with a Center which is not mentioned in the embedding DS.

[9] There were two separate occurrences of both **nullINIT** and **contSAME** in the database. Data from both occurrences of these targets will be included in the discussion of results in Chapters 6 and 7.

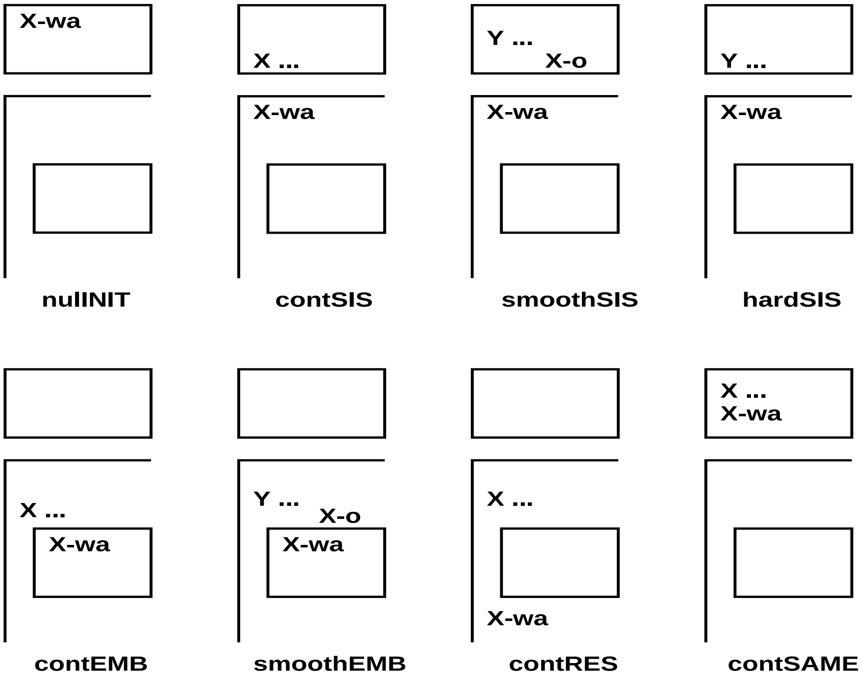

Figure 5.2: Schematic representations of the local focus target configurations summarized in Table 5.3.

5.3 Data collection

5.3.1 Speakers

The data presented in this thesis were collected from four native speakers of Standard Japanese: the variety spoken in and around Tokyo in the Eastern (Kantô) region of Japan. All speakers are female, ages 20-30, and were university students living in New York City at the time. None of the speakers were involved in the database construction, nor were any aware of the hypotheses being tested in the experiment. Table 5.4 summarizes the background of the four speakers.

5.3.2 Recording procedure

Recordings were made over a period of 5–6 (non-consecutive) days in an Industrial Acoustics Corp. sound-attenuated booth at the City University of New York (CUNY)

spkr	sex	age	birthplace (years lived)	other locations of residence (years lived)
MM	F	21	Yokohama-shi (18)	USA (3)
SN	F	25	Tokyo-to (25)	USA (.5)
KN	F	28	Yokohama-shi (23)	USA (5)
KF	F	30	Yamanashi-ken (19)	Shizuoka (6), Osaka (1), USA (3)

Table 5.4: Background of native speaker participants.

Linguistics Laboratory.[10] Each recording session on a given day lasted approximately 2 hours. The data were collected on DAT tape using a TASCAM (TEAC) DA-P1 portable DAT recorder and a Shure SM-10A directional head-mounted microphone, and were later digitally transferred to an SGI Workstation for analysis.

The nine discourses were printed on separate pages, and the order of the pages was randomized for each recording. In a given 2-hour session, the speaker read the set of nine discourses twice: once at the start, and once at the end of the session. The speaker recorded data for other experiments (not reported here) during the interval between reading these two discourse sets. All speakers were paid for their participation.[11]

Written instructions were presented (in Japanese orthography) on the first day of the experiment. The speaker was instructed to read each recipe as if she were instructing a new cook. She was told that the recordings would be played to the new cook on a future date, and the cook would need to answer a number of questions about the content and preparation of the recipes. The speaker was asked to read the discourses as fluently as possible, at a natural speed, with a loudness as if speaking to a person sitting beside her. A total of 9–10 repetitions of each discourse were elicited from each speaker in this way.

5.4 Data analysis

Recordings were transferred from DAT tape to an SGI (UNIX) workstation, and tagging and measurements were made using the Entropic Research Laboratories

[10] I am grateful to Di Bradley and Janet Fodor for allowing me to use the facilities at CUNY.

[11] The data collection was funded by a Graduate Student Alumni Research (GSARA) award from the Ohio State University Graduate School, and also by the OSU Department of Linguistics.

ESPS/Waves+ speech analysis software. The location of the high F0 target of the H* accent peak, as well as the intonational context surrounding each target NP, were tagged following the Japanese ToBI labeling scheme described in Section 2.1 and [Ven95].[12] The analysis presented in this thesis is restricted only to the variation of pitch range as measured by peak height on the targets, and possible systematic effects on other points in the contour (e.g. the trailing +L after the accent or the L% boundary, etc.) are not investigated here. Also, all targets involved in a speech disfluency or repair are excluded from the analysis.[13]

5.4.1 Measuring pitch range

The peak height of the target is taken to be a direct indicator of the pitch range of the phrase. In the model of pitch range and tone scaling adopted here (see Section 2.3), the range is defined as the tonal space bounded by the reference line and the topline. In implementations of this model for speech synthesis applications, the reference line is assigned a speaker-specific value, and remains constant throughout a sentence or discourse (e.g. [PB88, Spr98]). Therefore, given a constant lower bound, pitch range variation across phrases is directly reflected by changes in the upper bound: the topline. In these experimental data, this topline can be measured directly, since all target peaks are due to the H*+L lexical accent, which are scaled right at the topline according to this model of tone scaling.

While the peak F0 is taken to be a direct measure of the range topline, data presented in this thesis do not distinguish the type of topline at issue. That is, in a hierarchical model of pitch range such as that assumed here, the local topline of an accentual phrase (AP) can vary independently of the intonation phrase (IP) topline, as described in Section 2.3.3. For example, in cases of pragmatic focus on a specific lexical item, the local range can be expanded to cue this prominence. In such cases, the local range on other APs within the IP could remain as is (i.e. as predicted by phonological operations on the original IP topline) or could vary independently. In this study, all targets form their own accentual phrase which are all also IP-initial.

[12] Section 6.1 below will describe a case in which two of the speakers produced one of the target phrases as unaccented, characterized by a phrasal H- instead of an accent H*. In this case, the height of the 'peak' was taken to be the F0 value at the end of the initial rise (following the J_ToBI Guidelines). However, this measurement cannot be directly compared with the H* measurements, due to differences in the tone scaling of H- and H*, as described in Section 2.3.1. I will discuss this case in more detail below.

[13] When the disfluency occurred on or as a result of a mispronunciation of the target itself, no measurement is used. However, if the disfluency occurred after the speaker uttered the target (i.e. as a result of a mispronunciation of a word following the target), and the target phrase was repeated, the first occurrence (before the disfluency) of the target is considered to be not involved in the disfluency and so it is used in the analysis.

With only the measure of peak height on this AP, it is difficult to determine whether this height reflects the topline of the entire IP, or just the local range of the target AP (see the ambiguity noted in the description of Figure 2.4). In other words, any observed variation in pitch range of the target may be a result of a local range expansion/compression on the target only, or it may be due to a variation of the overall IP range. It is plausible that some discourse factors (such as the hierarchy of discourse segment intentions) affect the range of the whole IP, while other factors (such as local attentional focus) affect the range of the Centered entity only. But this is an empirical question. In order to investigate the relation between IP and AP toplines, one would need to examine the height of other AP peaks in the IP as well, in addition to this first target AP. This is an important issue, especially for the implementation of discourse rules in TTS or CTS systems, but I unfortunately will have to leave it for a future study.

5.4.2 Difference-from-mean dependent measure

The dependent measure examined in this thesis is not the raw F0 value of the topline as measured on the target peak. Rather, it is a 'normalized' F0 measure: the difference from the discourse mean. Since recordings were made for all speakers over a stretch of 5–6 days, and a stretch of 2 hours within each day, it is possible that the overall range that the speakers used for each discourse could vary considerably across and within recording sessions. For example, a speaker could use an expanded lively range at the beginning of a day's session, or could have a 'bad day' where her overall range is compressed. Such effects on the range of the target phrases are taken to be random, and I have attempted to factor them out by normalizing each target measure by the mean peak value of a given repetition of the entire discourse in which that target is situated.

In order to calculate the 'discourse mean', the high F0 of all accented /a/-peaks in the discourse were first tagged. These peaks include the discourse targets themselves, as well as all other phrases in the discourse meeting these criteria: they are all IP-initial nouns containing the accented low vowel /a/, and are not involved in a disfluency or repair. The mean of these /a/-peaks was then calculated for a given discourse repetition: that is, one production of a single discourse in a given recording set. This mean measure is taken to be the value (in hertz) of a 'default' IP topline, which can be used in a TTS implementation of that speaker's discourse. Then, the value of each target was subtracted from this overall discourse mean (of a given repetition), producing a difference-from-mean measure (also in hertz) which is reflective of the difference in pitch range of the target phrase from the default topline. It is this difference measure which is used as the dependent variable in the data analysis reported in this thesis. The measure abstracts away from random variation due to recording variability, emphasizes the fact that it is the relative value of the target peak with respect to the discourse in which it is situated that is important, and enables comparison of targets which were elicited in different discourses on different days. A positive difference measure indicates that the target is realized with a higher (expanded) range than predicted by the default topline, and a negative measure indicates that the target is realized in a lower (more compressed) range.

5.4.3 Statistical comparisons

Non-parametric statistical tests are used to judge the significance of the difference measure of a given target from the discourse mean (the 'default' topline), and also to test differences among the various target groups. Non-parametric statistics are chosen because they require fewer assumptions about the sample populations. That is, non-parametric methods do not require that the populations be normally distributed, nor that they have equal variances (e.g. [Gib93, WFH86], etc.). Two non-parametric tests, the Wilcoxon signed rank test and the Mann-Whitney U test (aka. Wilcoxon rank sum test), are used in the analysis presented in Chapters 6 and 7.[14]

The Wilcoxon signed rank test is used to test the significance of the difference measures of a given target group from the discourse mean. Like a sign test, the signed rank test considers the distribution of difference measures above and below the mean (diff=0). However, in addition, this test also incorporates information about the relative magnitudes of the differences in the test of significance. The Wilcoxon signed rank tests reported in this thesis are all one-tailed if the predicted prominence of the target is in one direction, and the tests are two tailed if the prediction is undetermined, or if there are conflicting predictions. Two-tailed tests will be marked with an asterisk '*'.

The second test used in the data analysis is the Mann-Whitney U test, also known as the Wilcoxon rank sum test. In this test, two samples are compared to one another. The dependent measures (i.e. the difference measures) are pooled then ranked according to their relative magnitudes. Significance is determined by comparing the sums of the rankings of the two groups.

The 5% significance level (the probability of rejecting the null hypothesis when it is true (Type I error)) is used for all statistical tests reported here. Significance at this level is marked by + or − signs in the results tables, indicating that the difference measure is significantly above or below the discourse mean, respectively.[15] The tag '(m)' marks cases in which the test is marginally significant at this level. In addition, significance at the 1% level is also marked by ++ or − −. In cases of multiple pairwise comparisons using the Mann-Whitney U test, the alpha (significance) level must be

[14] Both tests were implemented by hand using the MathSoft Splus statistics/graphics package, following the description provided in [Gib93].

[15] The + and − symbols are also used in the pairwise comparisons to show the direction of the effect — i.e. the height of the second target type relative to the first.

adjusted in order to correct for possible Type I errors due to multiple comparisons. The Bonferroni correction used is: α/J, where J is the number of pairwise comparisons. That is, setting the alpha level of each test to α/J (.05/J or .01/J here) assures that the probability of making at least one Type I error amongst all the pairwise comparisons is no larger than .05 (or .01). Such a correction yields α=.005 for significance at the 1% level and α=.025 for significance at the 5% level for 2 comparisons, and α=.003 (1% level) and α=.016 (5% level) for 3 comparisons. In the results tables of the target comparisons, the +/– signs take this correction into consideration.

5.4.4 Defining 'intonational prominence' in Japanese

Taking the difference-from-mean as the dependent measure, how can this be used to judge intonational prominence in Japanese? In English, intonational prominence is cued by the presence of a pitch accent: accented words are prominent, while unaccented words are not (see [Ayers96] for a psycholinguistic study of the perception of intonational prominence). Accented vs. unaccented is a categorical distinction, and this phonological contrast is encoded by symbolic labels in the ToBI labeling scheme. The categorical nature of accent in English lends itself quite nicely to studies which investigate the relation between discourse and intonation. For example, many studies use tallies of accented/unaccented words, and relate these raw numbers or 'percent accented' measures to the various discourse structures (e.g. [Brown83, Terk84, Naka97a], etc.).

However, in contrast to English, Japanese seems to use pitch range as a means to cue intonational prominence, which is by definition a continuous measure. How can such a continuum be divided into 'prominent' vs. 'non-prominent' ranges, for purposes of describing the intonation-discourse mapping?[16] For the purposes of the current analysis, I employ the sign (+ or –) of the difference-from-mean measure (as judged by a Wilcoxon signed rank test) as the main indicator of intonational prominence in Japanese: a significant positive difference indicates marked intonational *prominence*, and a significant negative difference indicates marked intonational *non-prominence*. That is, for implementation of discourse effects in speech synthesis systems, it is important to identify discourse configurations in which the phrasal pitch range differs markedly from the predicted 'default' topline (as generated by the algorithm based on phonological, syntactic, and other factors). Defining

[16] Of course, it is possible that there is no need to 'categorize' this continuous variable into a binary distinction at all. For example, even in English, accentual prominence can be further categorized into pre-nuclear vs. even more prominent nuclear accents (see e.g. [Ayers96] for discussion of the perception of this distinction), and the continuous pitch range variation on these accents may influence the degree of perceived prominence as well. However, to facilitate analysis here (and in other studies), such categorization is useful.

prominent/non-prominent in this way is useful for such implementation. In addition to these marked prominent/non-prominent cases, I also consider cases which are not significantly higher nor lower than the mean to be intonationally *non-prominent*, though not markedly so. These are cases in which there is no discourse effect, and thus no reason to vary the topline from the default in a synthesis system. In addition, for many comparisons, it is useful to gauge the degree of difference among target types as well. Therefore, in the data analysis I also compare target groups with one another (using a Mann-Whitney U test) to judge *relative* prominence.

CHAPTER 6
CUES TO DISCOURSE STRUCTURE AND GLOBAL ATTENTIONAL SALIENCE

Studies have shown that, in many languages, intonation can be used by speakers to cue the linguistic structuring of discourse, as well as to mark the salience of entities in the attentional state. This chapter examines the effects that both intention-based discourse structure and global attentional salience have on intonational prominence, namely pitch range variation, in a read Japanese database.

Results from this portion of the data analysis show that both discourse structuring and global attentional salience have an influence on the intonational realization of referring expressions in Japanese. There is a tendency for DS-initial phrases to be realized in a higher pitch range, and for final phrases to be realized with a lower pitch range. In addition, a majority of speakers mark a `new` (first mention) entity with an increased range, in relation to an entity whose antecedent is in the same DS (`same`). Speakers also can use intonation to cue the current global salience of entities mentioned previously in the discourse: an entity whose antecedent is in the same segment before an intervening embedded DS (`sameRES`) is non-prominent, while an entity whose antecedent was previously popped from the focus stack (`non-adj`) tends to be prominent. These results are presented in detail in the following sections.

6.1 Discourse structure effects

Based on the review in Chapter 4 of previous studies which describe the intonation-discourse mapping in both Japanese and other languages, the following hypotheses can be made about the predicted patterning of pitch range on the discourse structure targets. Table 6.1 summarizes the discourse features of each of these targets, and provides a schematization (using upward and downward arrows) to show the hypothesized prominence predictions based on these two discourse factors.[1]

[1] For ease of description, the factors hypothesized to affect intonation are enumerated separately from one another, though in actuality it is probable that speakers' behavior in spoken discourse will represent a combination of factors. In the schematizations of the hypotheses, an upward arrow marks a prediction of intonational prominence, while a downward arrow marks a prediction of intonational non-prominence. The lack of an arrow means that no prediction is made either way.

HYPOTHESIS 1 — DISCOURSE (INTENTIONAL) STRUCTURE: The prominence of a referring expression is influenced by its position in the discourse segment. DS-initial targets are predicted to have a higher pitch range, while DS-final targets are predicted to have a lower range.

HYPOTHESIS 2 — GLOBAL ATTENTIONAL SALIENCE: The prominence of a referring expression is influenced by its status in the global attention of the discourse participants. Entities which are globally salient are predicted to be realized with non-prominent intonation, while those which are not salient in the discourse are predicted to be realized with prominent intonation.

	1	2
	DS-internal	globally
target	position	salient
`init`	DS-initial ⇑	no ⇑
`med`	DS-medial	no ⇑
`fin`	DS-final ⇓	no ⇑

Table 6.1: Summary of discourse features and prominence predictions for discourse structure targets.

The discourse structure targets are NP-*o* direct object (non-Centered) phrases located in initial (`init`), medial (`med`), or final (`fin`) position of DS4. Based on previous studies of discourse structure effects on prominence in Japanese and other languages, described in Section 4.1, `init` is predicted to be intonationally prominent, while `fin` is predicted to be non-prominent (e.g. [Leh75, Yule80, Silv87, GH92, SG94, Ven96, VS96, HN96], etc.). The patterning of the DS-medial target is not as clear as that is targets in initial or final positions, so the prediction for `med` is left unspecified in the first effects column of the table. In terms of global attentional salience, all three of the discourse structure targets are the first mention in the current segment as well as in the discourse. Therefore, they are predicted to be intonationally prominent, based on a number of previous studies in English and other languages described in Section 4.2 (e.g. [Hal67, Brown83, Terk84, DH88, Naka97a, Naka97b], etc.). However, it is important to reiterate that although this effect has been observed in Japanese as well [Sugi96, HSOF94, SH95, HSK96, VS96], the effect is not as robust as in English.

Results show that, in this subset of the data, speakers adopted two different pronunciations of the discourse structure targets (all discourse structure targets are the same lexical item *mana'ita-o* 'cutting board-OBJ'). Speakers SN and KN consistently pronounced the target as intended: *mana'ita-o* with a lexical accent on the second mora. In contrast, speakers MM and KF consistently pronounced the word as unaccented: *manaita-o.* Both of these variants are apparently acceptable in Tokyo Japanese. However, due to this pronunciation variation, it becomes more complicated to interpret the results. That is, while the H* of the accented *mana'ita-o* is scaled right at the topline and therefore directly reflects the pitch range of the phrase, the phrasal H- of the unaccented *manaita-o* is scaled slightly below the topline, as described in Section 2.3.1. Since the discourse mean was calculated considering only accented /a/-peaks (see Section 5.4.2), it represents a 'default' topline value which the accented H* peaks are predicted to rest on, in the absence of any discourse (or other) effects. Therefore, the difference-from-mean measure examined here is predicted to be zero for the accented productions, but the unaccented productions are predicted to have negative differences (all else equal). That is, the unaccented cases are predicted to be realized slightly below this default topline, based on the tone scaling model outlined in Section 2.3.1. This potential confound due to differing speaker pronunciations should be kept in mind when reviewing the results for these targets.[2]

Table 6.2 summarizes the results for the discourse structure targets for all speakers. The table gives the number of tokens analyzed for each target, the median difference measure for that class, the p-value calculated by a Wilcoxon signed rank test (described in Section 5.4.3), and + or – symbols which summarize the statistical significance and direction of the effect.[3]

The results for speakers KF and MM are somewhat mixed. Speaker KF produced all discourse structure targets in a pitch range which is lower than the 'default' discourse mean. This pattern might be accounted for by the fact that the phrasal H- tones of her unaccented productions are scaled below the topline. Apparently any potential effects of DS-initiality or first mention of these targets are not robust enough to reverse this trend. In contrast, speaker MM, who also produced unaccented expressions in this data subset, shows an effect of DS-initiality. Her `init` target is

[2] All other targets in this thesis were pronounced with accented expressions, as intended.

[3] Here and elsewhere in this thesis, two-tailed tests are marked by an asterisk '*' next to the p-value. Those p-values without such marking are based on one-tailed tests. In this subset of the data, tests on the `fin` target for speakers SN and KN are two-tailed because this target has contrasting prominence predictions. However, for speakers MM and KF, who produced the target as unaccented, a majority of the hypotheses predict that the `fin` target will be realized below the discourse mean. Therefore, the tests on `fin` for these speakers are one-tailed. As for the `init` target, the tests are two-tailed for MM and KF, due to the conflict between the prediction of prominence in DS-initial position and the relatively lower F0 height of the unaccented phrasal H-.

	Speaker MM				Speaker SN			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
`init`	9	16	p=.012*	+	6	68	p=.016	+
`med`	9	7	p=.150		10	65	p=.001	++
`fin`	9	14	p=.248		8	15	p=.196*	

	Speaker KN				Speaker KF			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
`init`	7	20	p=.008	++	10	-43	p=.008*	– –
`med`	9	35	p=.002	++	10	-23	p=.001	– –
`fin`	9	21	p=.004*	++	9	-33	p=.002	– –

Table 6.2: Summary of results for the discourse structure targets.

realized with a pitch range above that of the (accented) discourse mean. The other speakers (SN and KN) both produced accented expressions, so their data are directly comparable to the discourse mean. However, the results are somewhat mixed for these speakers as well: SN realizes both DS-initial and medial targets with prominent intonation, while for KN, all of the discourse structure targets are prominent, including the `fin` target. It is possible that for this speaker, the effect of first mention of these targets overrides any DS-position effect, though further conclusions about the relative influences of these factors should be investigated in further experimentation using a more balanced design.

Comparison of the ranking of target median values for each speaker also show mixed results; only speakers MM and SN show a tendency for DS-initial targets to be realized in a higher pitch range than the other positions. However, this trend does not reach significance for either speaker. Perhaps if more data were collected for each condition, the results may become more robust.[4] In addition to this DS-initial effect, for three speakers (SN, KN and KF), the ranking of the `med` target median is higher than that for the `fin` target, suggesting a tendency for realizing DS-final targets in a lower pitch range. This trend reaches significance only for speaker SN (Mann-Whitney U test, p=.013).

[4] In future studies, it may also be useful to examine the initial/medial/final contrast in other discourse segments as well. A comparison of DS-initial targets in the other segments shows that the median of the DS4-initial target is ranked as one of the two lowest for three of the speakers (MM, KN and KF), suggesting that there may be another unknown factor which is causing a lowering of pitch range in this case. Therefore, future examination of targets in many different DS may yield a more robust effect of the trend observed in these data.

In sum, the data from this subset suggest that discourse position, as defined by intention-based segmentation and reinforced by 'paragraphing', influences the choice of pitch range on target expressions to some extent. Segment-initial phrases tend to be intonationally prominent for two of the speakers (MM and SN), while DS-final phrases tend to be realized by non-prominent intonation (SN, KN and KF). The effect of discourse structuring on intonational realization is most clearly observed in speaker SN's data.

6.2 Global salience: 'Given' vs. 'new'

The global focus targets can be used to examine the effect of global attentional salience on the intonational realization of referring expressions in Japanese. Previous studies have suggested that the discourse-'given' vs. 'new' distinction may be marked by pitch range variation in Japanese (see the review presented in Section 4.2.1), although the data bearing on this issue are limited. Here, I further investigate the influence of given/new as defined by global attentional salience. That is, I take 'new' to mean the first mention of an entity in the entire discourse (not globally salient), and 'given' to mean discourse-old information: that which has already been introduced into the focus stack at some previous point (see e.g. [Prince81] for a discussion of types of 'givenness'). In addition, the notion of 'given' is further constrained by the location of the entity with respect to the global focus stack. In this section, I consider a target entity to be 'given' if the antecedent was mentioned within the same focus space. I prefer to call this discourse configuration same to avoid confusion with the many definitions of the general term 'given'. Table 6.3 summarizes the discourse features and prominence predictions of the new and same targets.

	1	2
	DS-internal	globally
target	position	salient
new	DS-medial	no ⇑
same	DS-medial	yes ⇓

Table 6.3: Summary of discourse features and prominence predictions for new vs. same targets.

Both targets are NP-*o* (non-Centered) objects which are located in DS-medial position. There is no prominence prediction specified for DS-medial targets according to HYPOTH 1. The differentiating factor is only their global attentional salience: the `new` target is completely new to the discourse and as such is not considered globally salient, hence the prediction of prominence. In contrast, the antecedent of the `same` target is located in the current global focus space (though not in an immediately preceding utterance), and so the target is predicted to have non-prominent intonational marking according to HYPOTH 2.

All of the global focus targets were produced as accented referring expressions by all speakers, allowing for direct interpretation of the difference-from-mean measures. Table 6.4 summarizes the results for the `new` vs. `same` targets.[5]

	Speaker MM				Speaker SN			
target	*n*	mdn. diff(Hz)	p-val	sig.	*n*	mdn. diff(Hz)	p-val	sig.
`new`	9	19	p=.248		10	52	p=.001	++
`same`	9	25	p=.082		10	35	p=.010	++

	Speaker KN				Speaker KF			
target	*n*	mdn. diff(Hz)	p-val	sig.	*n*	mdn. diff(Hz)	p-val	sig.
`new`	9	35	p=.002	++	10	29	p=.001	++
`same`	9	22	p=.006	++	10	18	p=.001	++

Table 6.4: Summary of results for `new` vs. `same` targets.

Results indicate that the difference in global attentional salience among these two targets is not reflected by the difference-from-mean statistical tests. That is, for speaker MM, both `new` and `same` targets are realized with a pitch range not significantly different from the discourse mean, while for the other speakers (SN, KN and KF), both targets are realized significantly above the mean. In fact, the absolute values of both targets shown here are above the mean for all speakers (though for MM the large variance prevents the test from reaching significance). However, comparison of relative values of the target medians does indicate an effect of global salience. Speakers SN, KN and KF produce `new` targets which are realized with a higher range than their `same` targets. Only MM shows a reverse trend. In addition, the `new` > `same`

[5] All Wilcoxon signed rank tests performed on the global focus data are one-tailed, since the predicted prominence is in a single direction.

patterning of median ranking reaches significance (according to a Mann-Whitney U test) for speaker KF, and marginally for speaker SN, as summarized in Table 6.5. It is possible that future research which considers a larger number of tokens (n) in each category may yield more robust results.[6]

	Speaker MM		Speaker SN		Speaker KN		Speaker KF	
comparison	p-val	signif.	p-val	signif.	p-val	signif.	p-val	signif.
new,same	p=.365		p=.053	+(m)	p=.193		p=.045	+

Table 6.5: Summary of significant differences among new vs. same targets.

In sum, this subset of data shows an effect of 'given' vs. 'new' — or in terms of Grosz and Sidner's approach, of *global attentional salience* — on the relative pitch ranges of the targets. Three of the speakers (SN, KN and KF) chose to realize the new target with increased intonational prominence compared with the same target, and this difference reached significance for speakers SN and KF.

6.3 Global salience: Which 'given' entities are salient?

The preceding section described the intonational marking of 'given' information in which the antecedent occurs in the same global focus space as the target phrase. But what about other types of discourse-old information, in which the antecedent is located in other spaces on the focus stack? Comparison of the global focus targets non-adj, adj and sameRES further illuminates this question. Table 6.6 summarizes the discourse features and prominence predictions for these targets.

[6] Future research should also consider possible confounds of DS-internal structure on the realization of target heights. For example, in the case of the same target examined here, an independent discourse labeler (MU) in a post-hoc analysis judged this as a *possible* location of a DS-internal segment boundary (see discussion of this post-hoc analysis in Section 5.1.1). That is, she judged that the segment containing same could be broken down into two sub-segments, in which case the same target would be located initial to the second sub-segment. If the native speakers/readers perceived this DS (or 'paragraph') as having such internal structure, this may have caused speakers to produce the sub-segment-initial same target in a higher pitch range than expected. Future investigations using a less controversial DS-internal structure may find a more robust effect of 'given' vs. 'new'.

target	1 DS-internal position	2 globally salient
non-adj	DS-medial	no ⇑
adj	DS-medial	yes ⇓
sameRES	DS-medial	yes ⇓

Table 6.6: Summary of discourse features and prominence predictions for the non-adj, adj and sameRES targets.

As with new and same, these targets are all NP-*o* object (non-Centered) phrases which occur in DS-medial position. There is no prominence prediction specified for medial phrases by HYPOTH 1. In terms of their global attentional salience, the non-adj target is predicted to be prominent, based on studies showing that entities in global focus spaces which have been previously popped from the focus stack are no longer considered salient in the discourse. As such, these entities are marked by intonational prominence when re-introduced [VS96, Naka97a, Naka97b]. In contrast, entities in a focus space which has just been popped from the stack are still considered salient. This has been suggested by a number of studies examining the intonation-discourse mapping using the Grosz and Sidner model of attentional state [DH88, VS96, Naka97a, Naka97b], as well as studies which have described the surface form (full NP or pronoun) of referring expressions whose antecedents are in the previous focus space (e.g. [Walk98]). Therefore, based on these studies, the antecedent of the adj target is considered globally salient, and thus this target is predicted to have non-prominent intonation. As for the sameRES target, its antecedent is located within the same discourse segment: it is located in the first part of DS2, before the push of the embedded DS3. According to Grosz and Sidner's focus stack model, after the embedded segment is popped, the antecedent of sameRES is still in the focus stack (in fact, in the current re-opened focus space (FS2)), and so the entity is considered to be globally salient, and hence predicted to be intonationally non-prominent.

Analysis of these global focus targets shows an effect of global attentional salience on their intonational realization. Table 6.7 summarizes the data and results of Wilcoxon signed rank tests. These results show two interesting trends. First, the pitch range of non-adj targets is realized significantly above the discourse mean for three speakers (MM, SN and KF). This is consistent with the prediction that intonational prominence will mark the re-introduction of the entity into global focus.

A second observation is that the sameRES target is realized as non-prominent by all speakers: the pitch range is not significantly different from the 'default' topline prediction. This is also consistent with the predictions of Grosz and Sidner's focus

	Speaker MM				Speaker SN			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
non-adj	9	49	p=.027	+	10	51	p=.001	++
adj	9	13	p=.014	+	10	44	p=.003	++
sameRES	9	11	p=.125		10	8	p=.278	

	Speaker KN				Speaker KF			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
non-adj	9	-3	p=.285		10	15	p=.001	++
adj	9	-1	p=.410		10	-8	p=.500	
sameRES	9	5	p=.125		8	7	p=.055	

Table 6.7: Summary of results for non-adj, adj and sameRES targets.

stack model, in that the referent of the target is said to be globally salient, since the antecedent is located in a space still on the stack. The speakers' choice to refer to one of these entities using non-prominent intonation reflects this salience.

As for the adj target, the results are less clear. This target is predicted to be non-prominent, based on the data presented in previous studies which suggest that an entity contained within a just-popped focus space may remain salient in the attention of the discourse participants [DH88, VS96, Naka97a, Naka97b]. In the current study, this prediction is confirmed for only two of the speakers: KN and KF realize the adj target in a pitch range not significantly different from the discourse mean. In contrast, speakers MM and SN realize adj with prominent intonation. However, examination of the relative median rankings shows a clear contrast between the adj and non-adj discourse conditions. That is, the adj target is realized in a lower pitch range than non-adj by three of the speakers (MM, SN and KF), though this trend does not reach significance for any of the speakers.

In sum, the global attentional salience of an entity with respect to the global focus stack can influence its intonational realization. The data confirm results from previous studies by suggesting that speakers use prominent intonational marking on referring expressions which are completely new to the stack, or whose antecedent is in a non-adjacent popped focus space. In contrast, speakers tend to use a relatively less prominent intonational marking on expressions whose antecedent is in a just-popped space, or in the same space (even in the case of a re-opened space).

6.4 Local salience of the antecedent

The final issue investigated using the global focus targets is the possible effect that the local salience of an antecedent has on the intonational realization of a discourse target. Terken's analysis of pitch accent distribution in Dutch showed that referring expressions which are subsequent mentions of 'topics' tend to be realized with non-prominent intonation, in comparison with subsequent mentions of 'non-topics' [Terk84]. This is likely due to the fact that these subsequent mentions of 'topics' are locally salient Centered entities themselves, and this local salience is marked by non-prominent intonation, as described by Nakatani [Naka97a, Naka97b]. The effect of local salience on intonational prominence of targets in Japanese will be discussed in the next chapter. However, this brings up another interesting possibility: might the local salience of an antecedent have a direct effect on the realization of a non-Centered target (independent of the likelihood that a subsequent reference to a Centered entity will also be a Center)? That is, are some antecedents more salient than others, even when positioned in comparable discourse configurations? The additional global focus targets `non-adj(C)`, `adj(C)` and `sameRES(C)` were included in the database to address this question.

According to the model of attentional salience proposed by Grosz and Sidner, and studies of intonation-discourse mapping using this approach (e.g. [Naka97a, Naka97b], the local salience of an antecedent is not predicted to affect the intonational realization of a non-Centered entity. Nakatani describes the use of intonational prominence to cue a shift in local attention (Cb) to a new referent, but she does not discuss cases in which the new referent is not the Cb. Therefore, in cases where the target is not the Cb (as is the case with these global focus targets), the prediction is that the targets with Centered antecedents and the targets with non-Centered antecedents should behave similarly with respect to their intonational realization: both sets should pattern according to HYPOTH 2. Table 6.8 summarizes the results for both types of targets.[7]

The results in the table show that, contrary to the hypothesis, the local salience of the antecedent does have an effect on the intonational prominence of the target. In the case of the `sameRES(C)` condition, the target is realized with a pitch range significantly above the discourse mean for three of the speakers (MM, SN and KF), which is in sharp contrast to the `sameRES` target, which is realized by all speakers with non-prominent intonation. The `adj(C)` target also is produced with prominent intonation by three speakers (SN, KN and KF), only one of whom (SN) also had prominent intonation for `adj`. As for the `non-adj(C)` case, the target is realized as prominent by all speakers, which is consistent with the behavior in the non-Centered-antecedent case `non-adj`. In addition to the results from the Wilcoxon signed rank

[7] All Wilcoxon signed rank tests are one-tailed, since the prediction of prominence is in one direction, as described by HYPOTH 2.

	Speaker MM				Speaker SN			
target	*n*	mdn. diff(Hz)	p-val	sig.	*n*	mdn. diff(Hz)	p-val	sig.
`non-adj`	9	49	p=.027	+	10	51	p=.001	++
`non-adj(C)`	8	23	p=.004	++	10	47	p=.001	++
`adj`	9	13	p=.014	+	10	44	p=.003	++
`adj(C)`	8	11	p=.230		10	41	p=.010	++
`sameRES`	9	11	p=.125		10	8	p=.278	
`sameRES(C)`	9	25	p=.014	+	10	38	p=.019	+

	Speaker KN				Speaker KF			
target	*n*	mdn. diff(Hz)	p-val	sig.	*n*	mdn. diff(Hz)	p-val	sig.
`non-adj`	9	-3	p=.285		10	15	p=.001	++
`non-adj(C)`	9	11	p=.020	+	9	21	p=.004	++
`adj`	9	-1	p=.410		10	-8	p=.500	
`adj(C)`	9	18	p=.006	++	10	25	p=.001	++
`sameRES`	9	5	p=.125		8	7	p=.055	
`sameRES(C)`	9	12	p=.082		10	13	p=.001	++

Table 6.8: Summary of results for the two types of global focused targets: those with Centered antecedents, and those with non-Centered antecedents.

tests, examination of the relative ranking of the medians shows a clear tendency for the Centered-antecedent `(C)` targets to be realized with a higher pitch range than their non-Centered-antecedent counterparts. This trend holds for 8 of the 12 pairs shown in Table 6.8, though only KN's `non-adj`, KF's `adj` and SN's `sameRES` comparisons reach significance according to a Mann-Whitney U test (KN: p=.025, KF: p=.039, SN: p=.032).

These results suggest that the global salience of an entity with respect to the focus stack is just one of the factors which can contribute to its intonational realization. In addition, the local salience of the antecedent can also influence the speaker's choice of pitch range on a referring expression. In this case, reference to entities which were once locally salient Centers results in a increased prominence of the target. This is reminiscent of the phenomenon described by Nakatani [Naka97a, Naka97b], in which intonational prominence on a pronoun cues a shift in the local Center of attention back to something that was once in focus. The increase of intonational prominence on once-salient entities described here may be a related strategy that speakers can use to signal to the listener that "this is the same entity I was telling you about before that we need to talk about again". In this way, the prominence may mark the 'rehashing' of this crucial entity (a main ingredient in the recipe), much like the accentuation of **HE** to return to one of the main characters in Nakatani's data (see the monologue excerpt presented in Section 4.2.4). In these Japanese data, the prominence cues a return to a former Cb (albeit as a current non-Center), while in Nakatani's English data it cues the return to a former Cb as the current Center. In the next chapter, I will provide evidence that such rehashing of a Centered entity (as the current Cb) is also marked by intonational prominence in Japanese, similar to Nakatani's English data.

CHAPTER 7
CUES TO LOCAL ATTENTIONAL SALIENCE AND CENTERING RELATIONS

This chapter presents a detailed analysis of the effect of local attentional salience and Centering transitions on the intonational realization of referring expressions in the Japanese database. Results show that, in addition to effects of discourse structure and global salience, there are effects of the local attentional salience and of the relation that the Centered entity has with the immediately preceding discourse context on the realization of NP-*wa* referring expressions. In addition, two of the speakers show a clear interaction between Centering transition type and the hierarchical structure of the discourse. These results are presented in detail in the following sections.

The local focus targets examined in this chapter all function as the local Center of attention (the Cb) at the point in which they appear in the discourses, as described in Section 5.2.3. The hypotheses being tested here concern whether this local salience affects the intonational realization of the target, apart from any effects of global salience, and whether the type of Centering relation (transition) which links the target entity to the preceding discourse context also has an effect on its realization. Targets were placed in strategic locations in the discourses in order to investigate these questions. However, in addition to the systematic variation of local focus and Centering relations, targets may also differ in their position in the discourse segment, and/or their global attentional status. Therefore, in the discussion of the results in the sections below, I will reiterate the predicted patterning according to these other influencing factors (HYPOTH 1 and 2), in addition to outlining the predictions based on local salience influences.

7.1 Global vs. local salience

Experimental results reported in previous chapters of this thesis suggest that speakers can use intonational means to distinguish entities which are globally salient in the discourse from those which are newly introduced to the global focus space. Section 6.2 presented results from the current database which show that a non-Centered NP-*o* target whose antecedent is located in the same DS (`same`) tends to be realized with a lower pitch range than an otherwise comparable but `new` target. In addition, pilot data presented in Section 4.2.5 also suggest that referring expressions which are marked by the *wa* morphological topic marker are realized with a pitch range substantially lower than that predicted by phonological factors. However, from the pilot data alone, it is not possible to pinpoint the exact cause of this reduced pitch range. Is it because NP-*wa* expressions are typically *evoked* entities which are salient in the global focus stack, similar to globally salient non-Centered entities? Or is it their status as the local Center of attention (the most salient entity at that point in the discourse) which licenses the use of non-prominent intonation? In this chapter I will test the latter hypothesis:

> HYPOTHESIS 3 — LOCAL ATTENTIONAL SALIENCE: The prominence of a referring expression is influenced by its status in the local attention of the discourse participants. Centered entities are highly salient, and thus are predicted to be realized with non-prominent intonation (similar to the status of unaccented pronouns in English and other languages).

The data collected for the local focus NP-*wa* targets can be compared against the global focus NP-*o* targets to elucidate this issue. There is one pair of targets which can be compared: the global focus target `same` vs. the local focus target `contSAME`. Only these two targets are comparable in this database design because they are the only two which differ only in the local salience status, and no other factors. Table 7.1 summarizes the discourse features and hypothesized prominence predictions for each of these targets.

	1	2	3
	DS-internal	globally	local
target	position	salient	Center
`same` (NP-*o*)	DS-medial	yes ⇓	no
`contSAME` (NP-*wa*)	DS-medial	yes ⇓	yes ⇓

Table 7.1: Summary of discourse features and prominence predictions for non-Centered `same` vs. Centered `contSAME` targets.

The global NP-*o* target `same` is located in DS-medial position and is globally salient (due to an antecedent in the same DS) but not Centered. In contrast, the local focus NP-*wa* target is also located in DS-medial position and is globally salient for the same reason, but in this case, the target represents the local Center of attention.[1] Since DS-medial position is not predicted to be especially prominent or non-prominent (HYPOTH 1), the prediction for positional effects is left unspecified for both targets (although see footnote 3 below). The global attentional salience of both targets predicts that they will be realized with non-prominent intonation according to HYPOTH 2. The one hypothesis which predicts a difference in intonational realization between the two targets is HYPOTH 3: the target which is also locally salient (Centered) in the discourse is predicted to be realized with non-prominent intonation, in relation to the non-Centered target.

Table 7.2 gives a summary of median differences and Wilcoxon signed rank test results for each of these targets. Given that all predictions in Table 7.1 are in the same direction relative to the mean, these numbers are less informative than the numbers in Table 7.3 which gives results of pairwise comparisons using Mann-Whitney U tests.[2] Comparison of the median rankings of the `same` and `contSAME` targets shows that all speakers realize locally salient Centered discourse entities with a lower pitch range than their non-Centered globally salient counterparts. This difference reaches significance for both comparisons reported for speakers KN and KF, and for one of the comparisons reported for speaker SN (see Table 7.3).[3]

Other global vs. local targets in this study are not directly comparable, due to differences in either discourse structure, global salience, or Center transition type. For example, the `contSIS` DS-initial NP-*wa* target might be compared with the DS-initial *mana'ita-o* 'cutting board-OBJ' target (`initDS2`) in the same discourse position (see Section 5.2.1 for a description of the *mana'ita-o* target placement). However, in this case, `contSIS` is globally salient, while *mana'ita-o* is not, resulting in a confounded comparison. Likewise, the DS-initial `hardSIS` local focus NP-*wa* target could also be compared with the same DS-initial *mana'ita-o* (`initDS2`), since both are in the

[1] Note however that one target is a grammatical object, while the other is a grammatical subject/topic. This difference may influence the intonational realization, although the current study does not investigate the effect of grammatical role in Japanese independent from discourse salience and structural effects.

[2] Both `contSAME` and `nullINIT` (see Section 7.3) targets have two separate occurrences in the database. The two tokens will be referred to using (a) and (b). For the comparisons in Table 7.3, the Bonferroni correction yields α=.005 for significance at the 1% level, and α=.025 for significance at the 5% level, for 2 comparisons. In the table, the +/− signs take this correction into consideration.

[3] The discussion in Section 6.2 notes a possible confound of DS-internal structure on the `same` target, which was judged as *potentially* sub-segment-initial in a post-hoc discourse analysis by an independent labeler (MU). In the same analysis, both `contSAME` tokens were also judged as *potentially* sub-segment-initial. This DS-internal structure may or may not have been salient to the readers/speakers. However, even so, it doesn't represent a potential confound to this comparison, since both target types were judged similarly in the post-hoc analysis.

	Speaker MM				Speaker SN			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
`same`	9	25	p=.082		10	35	p=.010	++
`contSAME(a)`	9	7	p=.102		10	-9	p=.053	
`contSAME(b)`	9	9	p=.285		9	15	p=.024	+

	Speaker KN				Speaker KF			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
`same`	9	22	p=.006	++	10	18	p=.001	++
`contSAME(a)`	9	-28	p=.002	− −	10	-2	p=.348	
`contSAME(b)`	9	-19	p=.006	− −	9	-3	p=.102	

Table 7.2: Summary of results for non-Centered `same` vs. Centered `contSAME` targets.

	Speaker MM		Speaker SN		Speaker KN		Speaker KF	
comparison	p-val	signif.	p-val	signif.	p-val	signif.	p-val	signif.
`same,contSAME(a)`	p=.398		p=.002	++	p=.000	++	p=.001	++
`same,contSAME(b)`	p=.245		p=.218		p=.000	++	p=.000	++

Table 7.3: Summary of significant differences among non-Centered `same` vs. Centered `contSAME` targets.

same discourse position and are new mentions, though this comparison is also confounded by the fact that `hardSIS` cues a shift in attention, while *mana'ita-o* doesn't. Such shifts in the local Center of attention do affect the intonational realization of a referring expression, as I will describe in the next section. Thus, comparison of the intonational characteristics of specific discourse entities in spoken language databases should be done with great caution, due to the many effects that are known to influence intonational realization.

What Table 7.3 shows then is that, when it is possible to factor out the known effects and compare two phrases based on the difference in local salience alone, as with `same` and `contSAME` above, speakers choose a lower pitch range to realize the Centered entity, perhaps to signal its salience in the local discourse context. In sum, local salience does have a demonstratable effect, once global salience is taken into account.

7.2 Centering transitions to sister segments

The local salience of a discourse entity, in and of itself, is not the only factor that can influence its intonational realization. In addition, the relation of the Centered entity to the immediate preceding discourse context is also predicted to play an important role. Nakatani's investigation of pitch accent distribution, described in Section 4.2.4, showed that intonational prominence can be used to mark a shift in the local Center of attention to a new referent (as in the case of accented pronouns), while non-prominence tends to cue maintenance of a locally salient entity as discourse Center (as in the case of unaccented pronouns) [Naka97a, Naka97b]. This is summarized in the next hypothesis.

> HYPOTHESIS 4 — LOCAL SALIENCE RELATIONS: The prominence of a (Centered) referring expression is influenced by the relation that this Center has with the preceding local discourse context. A shift in the Center of attention is realized by prominent intonation, while a continuation of Center is realized by non-prominent intonation.

The local focus targets which involve different Centering transition types were designed to test this hypothesis. Here, I will discuss the *continue*, *smooth shift* and *'hard' shift* Center transitions, in which the target discourse segment is in a sister relation (SIS) to the previous segment (discussion of the other hierarchical discourse configuration (EMB) will be delayed until Sections 7.4.2 and 7.5 below). All three targets are in DS-initial position and are the local Center of attention (Cb), so they are predicted to pattern the same way by HYPOTH 1 and HYPOTH 3. However, the predictions by HYPOTH 2 and HYPOTH 4 contrast, as outlined in Table 7.4 below. In the hardSIS condition, an entity is introduced into the global focus space and the local Center of attention is also shifted. This target is predicted to be prominent by both influences. As for the smoothSIS target, it is already globally salient (by previous mention in a linearly-recent utterance), and as such is predicted to be non-prominent. However, this target serves to shift the Center of attention, and thus should be realized with prominent intonation. That is, there are conflicting predictions for smoothSIS. The contSIS target is predicted to be non-prominent by both hypotheses: it is both globally salient, and continues the Center from the previous discourse context.

Table 7.5 summarizes the results for these transition to sister segment (SIS) targets.[4] The raw data are plotted in Figure 7.1. The horizontal line drawn in each plotted dataset in the figure represents the median of that data, and the asterisks below each dataset mark a significant difference from the discourse mean, according to the Wilcoxon signed rank test (see the summary in Table 7.5).

[4] The Wilcoxon signed rank tests in Table 7.5 for the smoothSIS target are two-tailed, since the prominence predictions according to HYPOTH 1–4 are equally mixed. However, the overall prediction based on these hypotheses is for hardSIS to be prominent and contSIS to be non-prominent, so these are considered one-tailed tests.

	1	2	3	4
	DS-internal	globally	local	Centering
target	position	salient	Center	transition
`hardSIS`	DS-initial ⇑	no ⇑	yes ⇓	shift ⇑
`smoothSIS`	DS-initial ⇑	yes ⇓	yes ⇓	shift ⇑
`contSIS`	DS-initial ⇑	yes ⇓	yes ⇓	continue ⇓

Table 7.4: Summary of discourse features and prominence predictions for targets in the transition to sister DS discourse condition.

	Speaker MM				Speaker SN			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
`hardSIS`	9	40	p=.002	++	10	60	p=.001	++
`smoothSIS`	9	32	p=.004*	++	10	42	p=.002*	++
`contSIS`	9	7	p=.125		9	-23	p=.037	−

	Speaker KN				Speaker KF			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
`hardSIS`	9	34	p=.004	++	10	18	p=.001	++
`smoothSIS`	9	-4	p=.367*		10	1	p=1.00*	
`contSIS`	9	12	p=.020	+	10	-23	p=.014	−

Table 7.5: Summary of results for targets in the transition to sister DS discourse condition.

Table 7.5 and Figure 7.1 show a clear effect of Centering transition on the intonational realization of local focus targets: speakers tend to mark Center shifting by intonational prominence, and Center continuation by non-prominence. That is, the `hardSIS` target is realized with a pitch range significantly above the discourse mean, for all speakers. In contrast, three of the speakers (MM, SN and KF) realize the `contSIS` target with a range which is either equal to or lower than the 'default' prediction. As for the `smoothSIS` condition, speakers adopt different strategies: MM and SN cue this shift with intonational prominence, while KN and KF do not.

A comparison of the median rankings of the three target types shows that, for speakers MM, SN and KF, the `hardSIS` target is ranked highest, followed by `smoothSIS`, then `contSIS`. For speaker KN, the `contSIS` target is realized higher than would be expected according to this general trend. The results of Mann-Whitney U tests on these pairwise comparisons are given in Table 7.6.[5]

	Speaker MM		Speaker SN		Speaker KN		Speaker KF	
comparison	p-val	signif.	p-val	signif.	p-val	signif.	p-val	signif.
contSIS,smoothSIS	p=.057		p=.000	– –	p=.081		p=.026	
smoothSIS,hardSIS	p=.081		p=.018	–(m)	p=.031		p=.012	–
contSIS,hardSIS	p=.005	–	p=.000	– –	p=.001	– –	p=.001	– –

Table 7.6: Summary of significant differences among targets in the transition to sister DS discourse condition.

In sum, the data suggest that Japanese speakers cue the local salience relations in discourse by varying the pitch range of referring expressions. A 'hard' shift to a new Center of attention (Cb) results in an increased pitch range relative to the 'default' topline value, and relative to the other transition types. For two speakers, a smooth shift to a new Cb is also cued by increased range. The higher pitch range value of hardSIS relative to smoothSIS may be a result of not only a shift in local salience, but of hardSIS's new introduction of the entity into global focus as well. That is, the high range may due to the compounded effect of these separate factors. In contrast to the shift transitions, speakers tend to cue the continuation of the Cb with non-prominent intonation, by a pitch range at or below the 'default' topline. This marking of local salience relations is consistent with Nakatani's findings in English [Naka97a, Naka97b]. In her data, locally salient entities (pronouns) are marked by intonational non-prominence if they function to maintain the current Cb in local focus (analogous to contSIS here), while they can be marked by intonational prominence if they serve to shift the local Center of attention to a new discourse referent (analogous to hardSIS or smoothSIS here).

[5] The Bonferroni correction is used in Table 7.6, yielding α=.003 for significance at the 1% level, and α=.016 for significance at the 5% level, for 3 comparisons. In the table, '(m)' means that the comparison is marginally significant.

7.3 Discourse-initial null transition

Results presented in the previous section show that speakers mark 'hard' shifts in the Center of attention by intonational prominence. In such cases, local attention is shifted from an entity in the preceding discourse to a new referent (the NP-*wa* which has just been introduced). Does the intonational realization in such cases pattern similarly to the `nullINIT` configuration, in which local attention is 'initiated' absolute discourse-initially with the use of an NP-*wa* form? In the `nullINIT` cases, there is no previous utterance, and thus no $Cb(U_{n-1})$, so no Centering transition can be defined. However, `nullINIT` is similar to `hardSIS` in that the NP-*wa* entity is newly introduced as the Center. Therefore, the empirical question is whether the 'initiation' of a new Cb in `nullINIT` cases is marked by intonational prominence, as in the `hardSIS` configuration.

Table 7.7 summarizes the discourse features and prominence predictions for these targets. Both targets pattern the same with respect to their DS-internal position and global/local salience status, and so are not predicted to differ by HYPOTH 1, 2, or 3. However, the targets differ in the type of Centering transition, as mentioned above. The `hardSIS` target shifts the local Center of attention and as such is predicted to be prominent, while the `nullINIT` target 'initiates' the Center discourse-initially. There is no prediction of prominence for this case. In addition to Center transition, the two targets differ in one other regard: their position in the hierarchical discourse structure. The `nullINIT` target is absolute discourse-initial, while the `hardSIS` target is initial to a sister segment. As mentioned in Section 4.1, previous studies have shown that variations of pitch range can cue the hierarchical structure of a discourse (e.g. [HP86, Ayers94]). Specifically, large increases in range can mark major DS boundaries, and sub-segment boundaries can be marked by smaller increases. Based on this research, absolute discourse-initial position is predicted to be more prominent than sister DS-initial position. This hypothesis can be formulated as follows.

> HYPOTHESIS 1.1 — HIERARCHICAL DISCOURSE STRUCTURE: The prominence of a (DS-initial) referring expression is influenced by the position of the discourse segment in the hierarchical structure of the entire discourse. Targets positioned initial to segments at major discourse boundaries will have increased pitch range relative to those at sub-segment boundaries.

Table 7.8 gives the results for the `nullINIT` targets (there are two separate occurrences), and also reiterates results for the `hardSIS` target from the previous section. Results show that the majority of speakers (MM, SN and KN) realize both `nullINIT` targets with significantly increased pitch range, just as in the `hardSIS` case. A comparison of the relative ranking of the `nullINIT` and `hardSIS` medians shows that `nullINIT` tends to be realized in a lower relative pitch range than `hardSIS` (except for MM's anomalous `nullINIT(a)` token, which was significantly higher than her

	1	1.1	2	3	4
target	DS-internal position	hierarchical structure	globally salient	local Center	Centering transition
`nullINIT`	DS-initial ⇑	disc-initial ⇑	no ⇑	yes ⇓	null
`hardSIS`	DS-initial ⇑	sister	no ⇑	yes ⇓	shift ⇑

Table 7.7: Summary of discourse features and prominence predictions for `nullINIT` vs. `hardSIS` targets.

	Speaker MM				Speaker SN			
target	*n*	mdn. diff(Hz)	p-val	sig.	*n*	mdn. diff(Hz)	p-val	sig.
`nullINIT(a)`	9	92	p=.002	++	10	42	p=.001	++
`nullINIT(b)`	9	26	p=.002	++	9	26	p=.003	++
`hardSIS`	9	40	p=.002	++	10	60	p=.001	++

	Speaker KN				Speaker KF			
target	*n*	mdn. diff(Hz)	p-val	sig.	*n*	mdn. diff(Hz)	p-val	sig.
`nullINIT(a)`	9	23	p=.002	++	10	10	p=.024	+
`nullINIT(b)`	9	33	p=.002	++	9	-9	p=.037	–
`hardSIS`	9	34	p=.004	++	10	18	p=.001	++

Table 7.8: Summary of results for `nullINIT` vs. `hardSIS` targets.

`hardSIS`). Three of the other comparisons reach statistical significance according to Mann-Whitney U tests, as shown in Table 7.9.[6] [7]

Since a main difference between these two targets is the Centering transition (null vs. shift), the differences, where they exist, might suggest that Center shifting is marked by a greater increase in pitch range that Center 'initiation'. However, there are additional data that are relevant to this comparison. Measurements from

[6] The Bonferroni correction is used in Table 7.9, yielding α=.005 for significance at the 1% level, and α=.025 for significance at the 5% level, for 2 comparisons.

[7] Also note that since KN differed from the other speakers as to which of the two `nullINIT` cases was lower than `hardSIS`, it is difficult to explain the discrepancies in terms of the relative general availability of the lexical referents *mango* (`nullINIT(a)`) vs. *eggplant* (`nullINIT(b)`).

	Speaker MM		Speaker SN		Speaker KN		Speaker KF	
comparison	p-val	signif.	p-val	signif.	p-val	signif.	p-val	signif.
hardSIS,nullINIT(a)	p=.001	– –	p=.062		p=.016	+	p=.315	
hardSIS,nullINIT(b)	p=.057		p=.000	++	p=.302		p=.000	++

Table 7.9: Summary of significant differences among nullINIT vs. hardSIS targets.

mana’ita-o ‘cutting board-OBJ’ targets were also collected from all DS-initial positions in the discourses (see Section 5.2.1). A comparison of the median rankings of the *mana’ita-o* targets in the different DS-initial positions shows that the absolute discourse-initial *mana’ita-o* target is ranked lowest by three speakers (MM, SN and KN), and second lowest by KF. This suggests that the discourse-initial position may be marked as relatively non-prominent in general, rather than just being a tendency specific to the nullINIT targets. If this is the case, the behavior directly contradicts HYPOTH 1.1, in which absolute-initial position is predicted to have the highest pitch range.

One explanation for the speaker behavior could be that the prominence marking typically accompanying DS-initial phrases, which functions to cue the discourse segment edge, may not be as necessary in absolute-initial position. That is, in this position, it is already clear to the listener (at least in such a reading task) that the speaker is beginning the discourse. If this is the case, lowering due to this cause would apply across the board, not only lowering the non-Centered *mana’ita-o* targets in absolute-initial position, but also lowering the nullINIT NP-*wa* targets as well. Of course, this lowering is relative, since the nullINIT targets are still quite prominent with respect to the discourse mean. In such a scenario, at least two factors could be at work: the prominent marking of Center ‘initiation’ similar to Center shifting, as well as a slightly reduced range due to its discourse-initial position. However, this hypothesis is still tentative, and more data from future studies are required to fully understand the relative contributions of all the influencing factors.

7.4 Continue transitions

Results described in the two preceding sections demonstrate a clear effect of local attentional salience relations on the realization of target NPs in the database. Center shifting and ‘initiation’ is marked by intonational prominence, while Center continuation is marked by non-prominent intonation. Data presented in this section will show that, in addition to Center transition type, the hierarchical discourse structure in which the Centered target is placed also has an impact on the speaker's choice of pitch range on a referring expression.

There are four separate Center continuation target types contained in the current database, each positioned in a different location with respect to both the DS-internal as well as hierarchical discourse structuring. Table 7.10 summarizes the discourse features and prominence predictions for each of the continue targets.

	1	1.1	2	3	4
	DS-internal	hierarchical	globally	local	Centering
target	position	structure	salient	Center	transition
contSIS	DS-initial ⇑	sister ⇑	yes ⇓	yes ⇓	continue ⇓
contEMB	DS-initial ⇑	embedded	yes ⇓	yes ⇓	continue ⇓
contRES	DS-medial ⇑??	resumed ⇑??	yes ⇓	yes ⇓	continue ⇑??
contSAME	DS-medial	(medial)	yes ⇓	yes ⇓	continue ⇓

Table 7.10: Summary of discourse features and prominence predictions for the continue transition targets.

The contSAME target, discussed in Section 7.1, is a DS-medial NP-*wa* Center which continues the Cb within the same discourse segment. In contrast, the other cont targets continue the Cb across a DS boundary. In the case of contSIS, presented in Section 7.2, this target continues the Cb from a previous sister DS. The contEMB target is initial to an embedded DS, and continues the Cb from the immediately preceding embedding DS. And finally, contRES occurs immediately following the pop of the embedded DS, and continues the Cb from the last sentence of the hierarchically-recent outer segment. Table 7.11 gives a summary of the target medians and Wilcoxon signed rank tests of significance for these cont targets.[8] The raw data are also plotted in Figure 7.2. Two main effects observed in these data will be discussed in turn below: the effect of DS-internal structure, and the effect of hierarchical structure on the DS-initial Center continuation targets.

[8] The tests for contRES in Table 7.11 are two-tailed, since the prominence predictions for this target conflict.

	Speaker MM				Speaker SN			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
`contSIS`	9	7	p=.125		9	-23	p=.037	–
`contEMB`	9	38	p=.020	+	10	53	p=.001	++
`contRES`	9	53	p=.054*	+(m)	9	39	p=.002*	++
`contSAME(a)`	9	7	p=.102		10	-9	p=.053	–(m)
`contSAME(b)`	9	9	p=.285		9	15	p=.024	+

	Speaker KN				Speaker KF			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
`contSIS`	9	12	p=.020	+	10	-23	p=.014	–
`contEMB`	9	8	p=.125		10	14	p=.019	+
`contRES`	9	6	p=.040*	+	9	-4	p=.360*	
`contSAME(a)`	9	-28	p=.002	– –	10	-2	p=.348	
`contSAME(b)`	9	-19	p=.006	– –	9	-3	p=.102	

Table 7.11: Summary of results for the continue transition targets.

7.4.1 DS-internal structure

One important question that can be investigated using the Center continuation data is the extent to which DS-internal position affects the intonational realization of NP-*wa* targets. Results from Section 6.1 suggest that speakers MM and SN use an increased pitch range to mark DS-initial NP-*o* (non-Centered) targets, relative to other DS positions. In order to examine potential effects of DS-internal position for Centered NP-*wa* referring expressions, the DS-medial `contSAME` local focus target can be compared to the DS-initial `contSIS` target.[9] As outlined in Table 7.10, these two targets are comparable, in that they are both globally and locally salient, and continue the Center of attention. However, the targets contrast in their DS-internal position. Therefore, according to HYPOTH 1, the initial `contSIS` target is predicted to be realized with a higher pitch range than the medial `contSAME` target.

Table 7.11 and Figure 7.2 show that, for all speakers, the DS-medial `contSAME` targets are realized with non-prominent intonation: they fall either right around the discourse mean (speakers MM, SN, and KF) or significantly below it (speaker KN).[10]

[9] I use `contSIS` as the DS-initial target in this comparison, since it is predicted to be the most 'unmarked' of the three DS-initial continue targets.

[10] Note that these `contSAME` targets are realized using non-prominent intonation even though an independent labeler judged both tokens as being *possibly* sub-segment-initial in a post-hoc analysis.

Of course, this could be either because of the Centering transition (the contSAME targets continue the Center), or because of the discourse position (they are all medial). We can tease apart the explanations by comparing them to the contSIS targets: if the effect is due to the influence of the discourse structure on the pitch range of the target (by HYPOTH 1), then we would also predict that the DS-initial target (contSIS) would be significantly higher than the discourse mean, or at least higher than contSAME. Table 7.12 shows results of pairwise comparisons between contSAME and contSIS targets.[11]

	Speaker MM		Speaker SN		Speaker KN		Speaker KF	
comparison	p-val	signif.	p-val	signif.	p-val	signif.	p-val	signif.
contSIS,contSAME(a)	p=.500		p=.200		p=.000	++	p=.038	
contSIS,contSAME(b)	p=.398		p=.003	– –	p=.000	++	p=.248	

Table 7.12: Summary of significant differences among contSAME vs. contSIS targets.

An effect of DS-internal position in the predicted direction is observed only for speaker KN: her DS-initial contSIS target is significantly higher than the discourse mean, and higher than both DS-medial contSAME targets, a behavior which is consistent with HYPOTH 1. However, the other three speakers (MM, SN and KF) show a very different pattern. For these speakers, the DS-initial contSIS target is realized either at the same height, or even *lower* than the medial targets. For speaker SN, the difference between the initial vs. medial targets is statistically significant, as shown in Table 7.12. This marked reversal of the initial/medial patterning for MM and KF, and especially for speaker SN, is unexpected. It is not consistent with any of the hypotheses outlined in Table 7.10, which all predict contSIS to have either the same or greater prominence than contSAME. As I will show in Section 7.6 below, this marked lowering of the contSIS targets is a key element of the interaction between Centering transition and hierarchical discourse structure.

[11] The Bonferroni correction is used in Table 7.9, yielding α=.005 for significance at the 1% level, and α=.025 for significance at the 5% level, for 2 comparisons.

7.4.2 Hierarchical structure

The second important question that can be investigated using the Center continuation data is the extent to which the hierarchical discourse structure affects the intonational realization of the DS-initial NP-*wa* Center continuations. As outlined in Table 7.10, the `contSIS`, `contEMB` and `contRES` targets are all located in DS-initial position, are globally/locally salient, and continue the local Center of attention from the preceding sentence of a linearly-recent sister DS, of a linearly-recent embedding DS, or of a hierarchically-recent (but not linearly-recent) outer DS, respectively.

First, the patterning of `contRES` will be considered. This target is difficult to incorporate into the hypotheses presented in Table 7.10, due to its unique position in the discourse. It is technically medial to DS2, although it is placed right after the pop of embedded DS3. This large discourse juncture gives it a 'flavor' of initiality, resulting in the tentative '??' prediction of prominence in HYPOTH 1 AND 1.1. The antecedent of `contRES` is located in the same focus space (in the portion of DS2 preceding the embedded DS3), and so the entity referred to by `contRES` is considered globally salient, and thus predicted to be non-prominent by HYPOTH 2. As the local Center of attention, it is also predicted to be non-prominent (HYPOTH 3). However, even though it continues the Center of the hierarchically-recent Cb in the outer segment after the pop of embedded DS3, studies suggest that this configuration should actually be considered a Center shift, since the Center is shifted from the linearly-recent sentence (in DS3) back to the previous Cb before the embedding. This is the reason for the tentative prominence prediction in HYPOTH 4. The data for this target presented in Table 7.11 and Figure 7.2 show that `contRES` is indeed marked by intonational prominence for three speakers (MM, SN and KN), all of whose productions are significantly higher than the discourse mean. Only KF shows no difference from the 'default' topline. Therefore, this result suggests that Japanese speakers can choose to mark a discourse Center with prominent intonation, if that Center serves to shift attention back to a previous Center. In fact, this target type could have been called a 'shift' rather than a 'continue'. This behavior is consistent with Nakatani's account of a subclass of accented pronouns (such as the **HE** example, see Section 4.2.4) in her spontaneous narrative data [Naka97a, Naka97b]. However, Nakatani also notes that prominence marking on the pronoun in such cases is actually optional. In fact, in her monologue data, half of the pronouns used to continue the Center of the outer segment after the pop of an embedded DS are realized as accented (7/15), while the other half are unaccented (8/15). This leads her to conclude that "while there may be a tendency to signal with accent a global shift in centers ... it is clearly not necessary to do so" [Naka97a, p. 148]. The 'optional' nature of this marking may be the reason why KF shows no increased pitch range on the `contRES` target, unlike the other speakers. In addition to being consistent with Nakatani's findings, marking the return to a previously Centered entity with prominent intonation is also consistent with the results presented in Section 6.3 above, in which the 'rehashing' of a previous

Center as a non-Cb is also marked by an increase in pitch range, in comparison to cases in which the antecedent is a non-Centered entity.

Finally, the last comparison in this data subset is that of the `contSIS` vs. `contEMB` targets. As outlined in Table 7.10, in both cases, the target is a DS-initial globally/locally salient Center which is continued from an preceding linearly-recent Cb in the previous DS. As such, both are predicted to pattern similarly by HYPOTH 1, 2, 3 and 4. The only hypothesis which predicts a difference is HYPOTH 1.1, in which phrases initial to embedded DS are predicted to have a lower pitch range relative to those initial to sister segments. As discussed above, results for `contSIS` show that speakers MM, SN and KF realize this target with non-prominent intonation, with SN and KF's productions being significantly below the discourse mean. On the other hand, the `contEMB` target is realized using prominent intonation by these same three speakers: the majority of the tokens for this target being significantly *above* the discourse mean. The difference between the two targets reaches significance for SN and KF (Mann-Whitney U test, SN: p=.000, KF: p=.002). This result suggests that, for speakers MM, SN and KF, there is an effect of hierarchical discourse configuration on the pitch range of referring expressions which continue the local Center of attention from a linearly-recent sentence. However, this influence of hierarchical structure in actually *opposite* of that predicted by HYPOTH 1.1. Instead of the embedded DS-initial phrase being realized with a lower pitch range, it is realized by a higher range, relative to the sister DS-initial position. I discuss this interesting effect in more detail in Section 7.6 below, after reviewing the results from the `smoothSIS` and `smoothEMB` targets in the following section.

7.5 Smooth shift transitions

Another example of the effect of hierarchical structure on the intonational realization of NP-*wa* referring expressions in Japanese is seen in the smooth shift transitions. The database includes two types of smooth shifts: shifting to a Center in a sister DS (`smoothSIS`) and shifting to a Center in an embedded DS (`smoothEMB`). A summary of the prominence predictions for these two configurations is given in Table 7.13.

Both targets are globally and locally salient, and as such are predicted to be non-prominent by HYPOTH 2 and 3. However, HYPOTH 4 predicts that both targets will be prominent, due to the fact that they serve to shift the Center of attention to a new referent. Likewise, HYPOTH 1 predicts that they will be prominent, since both are located in DS-initial position, and HYPOTH 1.1 goes one step further, to predict that the `smoothSIS` target will be realized with a higher range than `smoothEMB` due to differences in the hierarchical discourse structure. Table 7.14 gives a summary of

target	1 DS-internal position	1.1 hierarchical structure	2 globally salient	3 local Center	4 Centering transition
smoothSIS	DS-initial ⇑	sister ⇑	yes ⇓	yes ⇓	shift ⇑
smoothEMB	DS-initial ⇑	embedded	yes ⇓	yes ⇓	shift ⇑

Table 7.13: Summary of discourse features and prominence predictions for the smooth shift transition targets.

	Speaker MM				Speaker SN			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
smoothSIS	9	32	p=.004*	++	10	42	p=.002*	++
smoothEMB	9	44	p=.004*	++	10	10	p=.232*	

	Speaker KN				Speaker KF			
target	n	mdn. diff(Hz)	p-val	sig.	n	mdn. diff(Hz)	p-val	sig.
smoothSIS	9	-4	p=.734*		10	1	p=1.00*	
smoothEMB	9	-8	p=.054*	–(m)	10	-17	p=.002*	– –

Table 7.14: Summary of results for the smooth shift transition targets.

the results for the smooth shift targets, and Figure 7.3 plots the raw data.[12] The result of a pairwise comparison of the two targets using a Mann-Whitney U test is given in Table 7.15.

Results indicate that speakers MM, SN and KF all show a significant effect of hierarchical structure on the smooth shift targets. However, the direction of the trend is not the same for all speakers. For speaker MM, the smoothSIS target is significantly less prominent than the smoothEMB target, mimicking her continue transition data, but contradicting HYPOTH 1.1. In contrast, speakers SN and KF realize the smoothSIS target with a significantly *higher* pitch range than smoothEMB, which is consistent with HYPOTH 1.1. KN shows the same trend, but the difference does not reach significance. This pattern observed for SN and KF is opposite to that seen in their continue transition targets: in the continue transitions, SN and KF realize EMB significantly higher than SIS, while in the smooth shift transitions, they realize

[12] In Table 7.14, the Wilcoxon signed rank tests are two-tailed, since the prominence predictions conflict.

	Speaker MM		Speaker SN		Speaker KN		Speaker KF	
comparison	p-val	signif.	p-val	signif.	p-val	signif.	p-val	signif.
`smoothSIS,smoothEMB`	p=.007	– –	p=.007	++	p=.170		p=.006	++

Table 7.15: Summary of significant differences among targets in the smooth shift transition condition.

`EMB` significantly lower than `SIS`. That is, there is an interaction between the Center transition type and the hierarchical discourse structure for these two speakers. The following section examines this in detail.

7.6 Interaction of Center transition with hierarchical structure

Results in the preceding sections suggest that the speaker's choice of pitch range on an NP-*wa* target is determined not only by the type of Centering transition involved in making that entity locally salient, but also by the position of the target in the hierarchical discourse structure. Moreover, for some speakers there is an interaction between these two factors. The plots in Figure 7.4 summarize this interaction. Transitions to a sister DS are shown by hollow circles, and transitions to an embedded DS are shown by filled circles.

The figure shows that speaker KN displays no interaction between transition type and hierarchical structure: the targets which continue or smooth shift the Center to an embedded DS are both realized 4 Hz below those which continue/shift to a sister DS, a difference which is not significant.

Speaker MM, on the other hand, realizes the embedded DS targets consistently *above* the sister DS targets. The `smoothEMB` target is significantly above `smoothSIS` by 12 Hz, while `contEMB` is realized 31 Hz above `contSIS` (though this latter difference is not significant, due to the large variance). This pattern suggests a possible interaction of transition type with hierarchical structure for MM, although in this case, the interaction does not reverse the direction of the difference between the sister and embedded structures.

In contrast to speakers KN and MM, speakers SN and KF show a clear interaction between transition type and hierarchical structure, as shown by the criss-cross in the plots. For these speakers, the embedded DS target is more prominent than the sister DS target *only* in the continue transition condition. For the smooth shift transition, the reverse is true. This interaction suggests a unique strategy adopted by these two speakers to integrate structure with local salience relations. In the sister DS condition (hollow circles), the pattern that these speakers exhibit is consistent with the predictions of HYPOTH 4. That is, a shift in Center results in a choice of pitch range on the referring expression which is higher than in a continuation of the Center. This result is consistent with Nakatani's observations of pitch accent on pronouns in English [Naka97a, Naka97b]. However, the effect is reversed in the embedded DS condition in these data. What is it about an embedded DS that might make smooth shifts *less* prominent than continuations?

One possible explanation is that, for these speakers, introduction of the target entity using a direct object NP-*o* expression creates a configuration where 'zooming in' on that entity by pushing an embedded space onto the focus stack is a likely next step. This would result in the use of a non-prominent expression to smooth shift the Center to an embedded DS. In these discourses, the embedded DS represents a cooking tip for successful completion of the recipe. In the case of the `smoothEMB` target examined in this database, the embedded segment is an elaboration of the type of cooking pot to be used for the stew:

> *... kyô-no zairyô-wa moyashi, gobô, nama shîtake, shungiku, naganegi, soshite hakusai-no kimuchi desu. gudakusan-ni narimasu kara, ôkime-no nabe-o yôi shite kudasai.*
>
> *nabe-wa* (`smoothEMB`) *tessei-no ôki-na atsude-no mono-ga yoi deshô. ichido atsuku nattara ato-wa ...*

> ... Today's ingredients will be bean sprouts, burdock root, fresh shitake mushrooms, chrysanthemum leaves, scallions, and cabbage kimchee. Since there are many ingredients, please prepare a large pot.
>
> The pot (`smoothEMB`) should be a large thick cast-iron one. After heating it once ...

The direct object *nabe-o* acts like a pivot. It sets up the entity which becomes the Center at the beginning of the embedded segment.[13] Speakers SN and KF mark such a shift with non-prominent intonation, perhaps to cue the listener that 'zooming in' on a direct object is a likely function for an embedded segment.

In contrast, when the Center continues into an embedded segment, the same speakers mark the continued Center with prominent intonation. In this case, the embedded DS also 'zooms in' on an entity in the previous utterance, but it is already locally salient. That is, here, the discourse elaborates on the subject *mango* and how specifically to remove the pit:

[13] Brennan [Bren95] and Turan [Turan98] describe the prevalent use of EXPLICIT OBJ → EXPLICIT SUBJ → (ZERO) PRONOUN to smooth-shift Centers within discourse segments in English and Turkish, respectively.

> *... kirei-ni aratte, sûpâ-no shîru-ga tsuite iru bâi-wa ato-ga nokoranai yô-ni teinei-ni hagashimasu. kawa-o tsuketa mama tate-ni hanbun-ni kitte, tane-o torinukimasu.*
>
> *mango-wa* (`contEMB`) *momo-no yô-ni chûô-ni ôki-na tane-ga arimasu kara, torinuku-no-ni-wa sukoshi kotsu-ga irimasu. mango-o kawa-o shita-ni shite ...*

> ... wash it clean, and carefully remove the supermarket sticker if there is one. Leave the skin on and cut it in half lengthwise, then take out the pit.
>
> Since mangos (`contEMB`) have large pits like peaches, you need a bit of technique to get them out. Hold the mango with skin side down ...

The target Center *mango-wa* which is continued into the embedded DS is marked by prominent intonation in the two speakers' data. This may be a strategy to cue the listener that the speaker is in fact still discussing the same entity (the mango), but she is changing the angle: she is 'zooming in' on the pit, whereas before she was talking about the mango as a whole. Thus, in contrast to the `smoothEMB` case, where the speaker uses non-prominent intonation to cue the elaboration of a direct object as a likely function of a embedded DS, in the `contEMB` case she must use prominent intonational marking to signal to the listener that she is giving additional information about an already Centered entity, but it is information of a different type. In other words, the speaker uses prominent intonation to cue the momentary shift in the viewpoint. This explanation can be formalized as the following hypothesis.

> HYPOTHESIS 5 — EMBEDDING FUNCTION (amendment to HYPOTH 4): The prominence of a (Centered) referring expression which is initial to an embedded segment is influenced by the type of local salience relation which the target entity has with the immediately preceding discourse context. Targets which 'zoom in' on a non-Centered object mentioned in the linearly-recent utterance may be marked with non-prominent intonation. In contrast, targets which continue the subject Center, but change the viewpoint of the information provided, may be marked with prominent intonation.

This choice of intonational marking in the embedded DS condition is opposite of that in the sister DS condition: in the `SIS` case, the speaker cues continued discussion of a referent by non-prominent intonation, and a shift in attention by relatively more prominent intonation, as predicted by previous studies (see HYPOTH 4). More experimental data which include Centered targets in embedded DS are needed in order to better understand this interesting interaction between local salience relation and hierarchical discourse structure.

7.7 Summary

This chapter has demonstrated that Japanese speakers can use intonational means to mark the status of a discourse referent with respect to the local attention of the discourse participants. Specifically, experimental speech production data are presented which show that speakers use pitch range variation to mark the salience of an entity at the local vs. global focus level, and also to mark the relation that entity has to the previous discourse context. In addition, some speakers show an interaction of this local Centering relation with the hierarchical structure of the discourse. Results suggest that speakers do not always adopt the same strategy in marking these discourse phenomena, though the inter-speaker variation does operate within the bounds of predictions made about the relationship between intonation, attentional salience relations, and discourse structure. Table 7.16 summarizes the significant differences from the discourse mean for all of the local focus targets presented in this chapter. Below, I will discuss how these results fit with the hypotheses outlined in this chapter, and also discuss how speaker strategies may vary.

transition type	MM	SN	KN	KF
nullINIT(a)	++	++	++	+
nullINIT(b)	++	++	++	–
hardSIS	++	++	++	++
smoothSIS	++	++		
smoothEMB	++		–	– –
contSIS		–	+	–
contEMB	+	++		+
contRES	+(m)	++	+	
contSAME(a)		–(m)	– –	
contSAME(b)		+	– –	

Table 7.16: Summary of significant differences from discourse mean for all local focus targets.

None of the hypotheses is able to account for all of the significant differences and speaker variation that is shown in Table 7.16. The local attentional salience hypothesis (HYPOTH 3) is able to account for the non-prominent marking of locally salient Centers (contSAME) in comparison to globally salient non-Centered entities (same), although the variation in prominence marking on other local Centers shown in the table suggests that there is more to the story. Likewise, the 'flat' discourse structure hypothesis (HYPOTH 1) is able to account for the non-prominence of the DS-medial Center (contSAME), although it does not predict the marked non-prominence in cases of Center continuation to a sister DS (contSIS) or smooth-shifting to an embedded DS (smoothEMB) seen in most speakers' productions. In addition, the hierarchical variant of the discourse structure hypothesis (HYPOTH 1.1) cannot account for all the effects of hierarchical structure observed in the data either.

The most promising hypothesis is the local salience relations hypothesis (HYPOTH 4, along with the HYPOTH 5 amendment for some speakers), and to some extent the global attentional salience hypothesis (HYPOTH 2). All speakers cue Center shift or 'initiation' (hardSIS and nullINIT) using prominent intonation, while using non-prominent intonation to cue Center continuation within a discourse segment (contSAME) and across a DS boundary to a sister segment (contSIS).[14] This behavior is predicted by both the local relations and global salience hypotheses. However, for the other target types, speaker strategies differ, and can mostly be described by one or the other of these two hypotheses. Each speaker's strategy is described in detail below.

- **Speaker KN:**

  In general, this speaker's behavior patterns with the global salience hypothesis (HYPOTH 2), in that she realizes newly introduced entities with prominent intonation, and most of the globally salient entities with non-prominent intonation. This is especially noticeable in the smooth shift condition, where she chooses non-prominent intonation to mark the Center shifts. Apparently, any effect of Center transition (shifting) is overridden by the global salience of these targets. However, this speaker does mark the contRES target as prominent, which is consistent with the local salience relations hypothesis (HYPOTH 4). In addition, her contSIS target is prominent, which cannot be accounted for at the current time.

- **Speaker MM:**

  This speaker's data are consistent with the local salience relations hypothesis (HYPOTH 4). She marks all Center shifting (including contRES and 'initiation') with prominent intonation, and marks Center continuation within the DS and to a sister DS using non-prominent intonation. Her prominent marking of contEMB suggests that the 'embedding function' amendment (HYPOTH 5) may also be relevant for this speaker.

[14] Only speaker KN realizes contSIS as prominent. This cannot be accounted for by any of the current proposals.

- **Speaker SN:**

  This speaker's behavior is a perfect example of a combination of the local salience relations hypothesis (HYPOTH 4) and the 'embedding function' amendment (HYPOTH 5).[15] Her Center shifts and 'initiation' are prominent, with the exception of the non-prominent `smoothEMB`, which is predicted by HYPOTH 5. Her Center continuations are non-prominent, as expected by HYPOTH 4, with the exception of the prominent `contEMB` accounted for by HYPOTH 5.

- **Speaker KF:**

  This speaker's behavior is the most difficult to explain. Like speaker SN, she shows a clear interaction between Centering transition and hierarchical discourse structure, so the HYPOTH 5 amendment is relevant here also. Her realization of most targets is consistent with either the global salience (HYPOTH 2) or local salience relations hypotheses (HYPOTH 4), although there are a few exceptions. The lack of prominence on `nullINIT` clearly contradicts HYPOTH 2, while the non-prominence of `contRES` contradicts HYPOTH 4. In addition, the non-prominence of `smoothSIS` appears to contradict HYPOTH 4, although it is realized with a high range in relation to the `smoothEMB` target.

In sum, this chapter presents a detailed analysis of the effect of local attentional salience relations and hierarchical discourse structure on the intonational prominence of the local focus Centers. Results show that both factors have an effect on the realization of target NPs, and for some speakers, there is an interaction between the two. Another important observation is that speaker strategies for marking prominence do differ, though the data suggest that this inter-speaker variation operates within the bounds of predictions made about the relationship between intonation, attentional salience relations, and discourse structure.

[15] Interestingly, this speaker was judged to have an 'excellent reading voice' by an independent native speaker before any of the data were analyzed (Yuki Hirose, personal communication, June 1999).

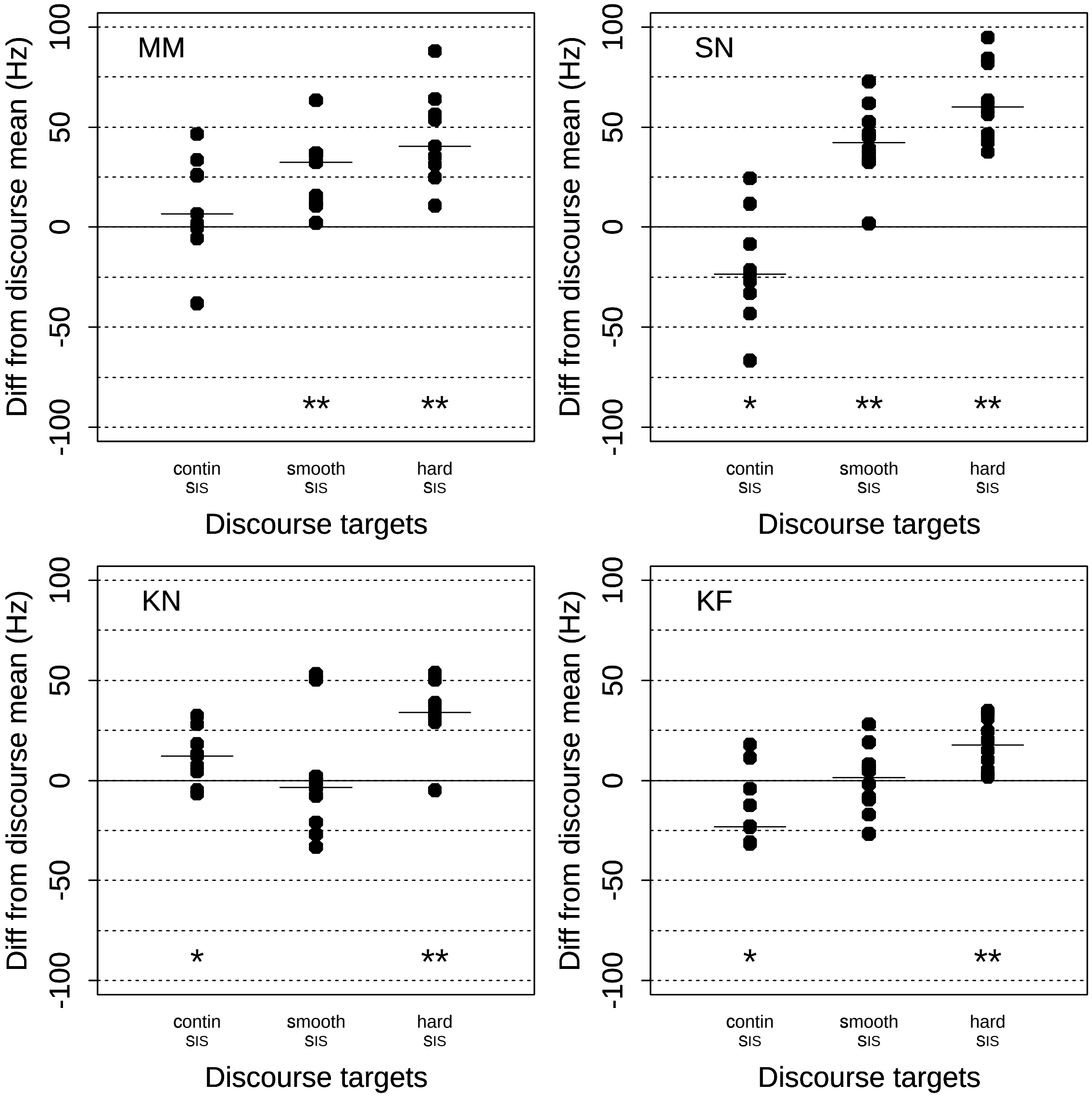

Figure 7.1: Difference measures for transitions to NP-*wa* in a sister segment.

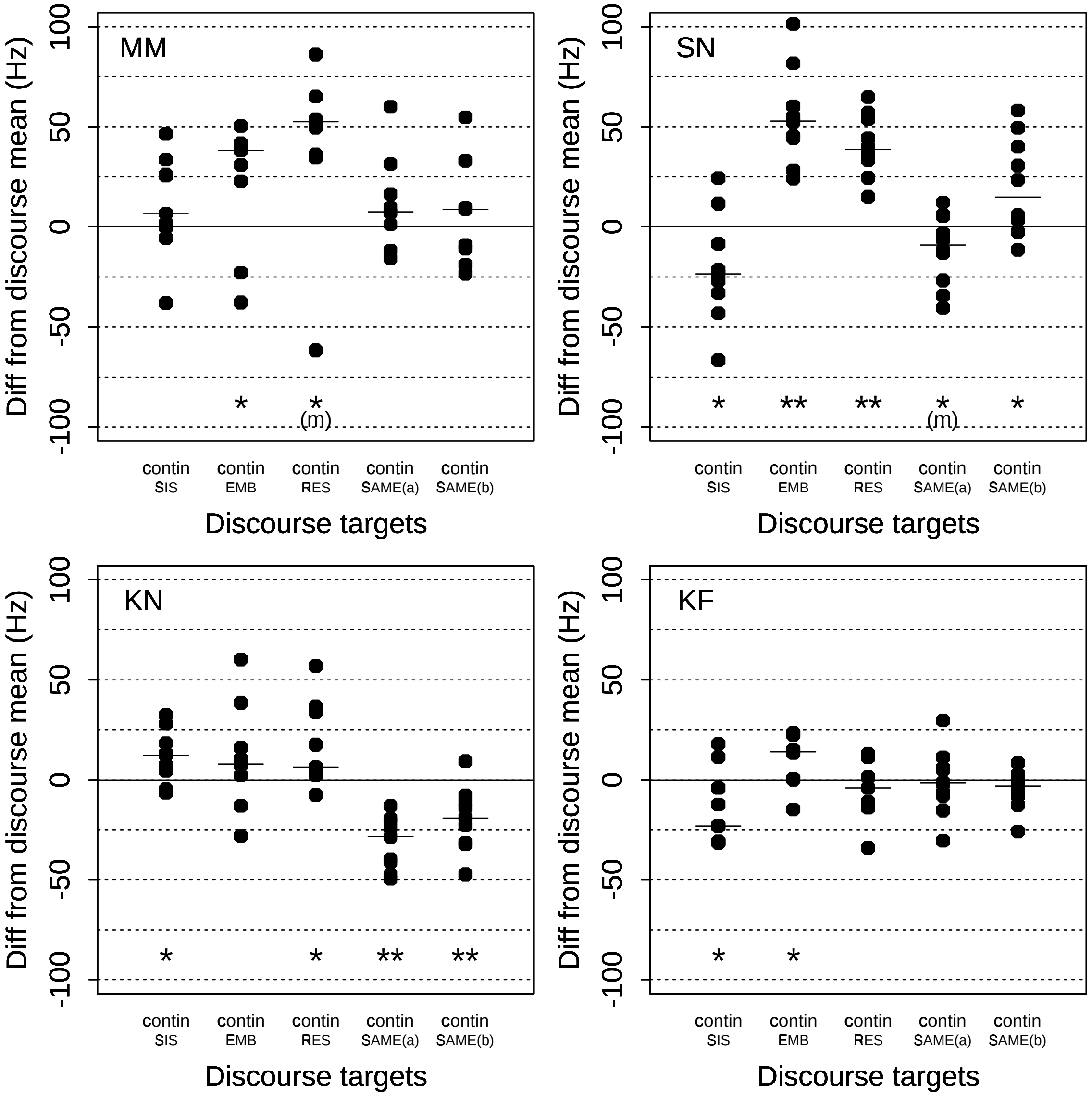

Figure 7.2: Difference measures for NP-*wa* in continue transitions.

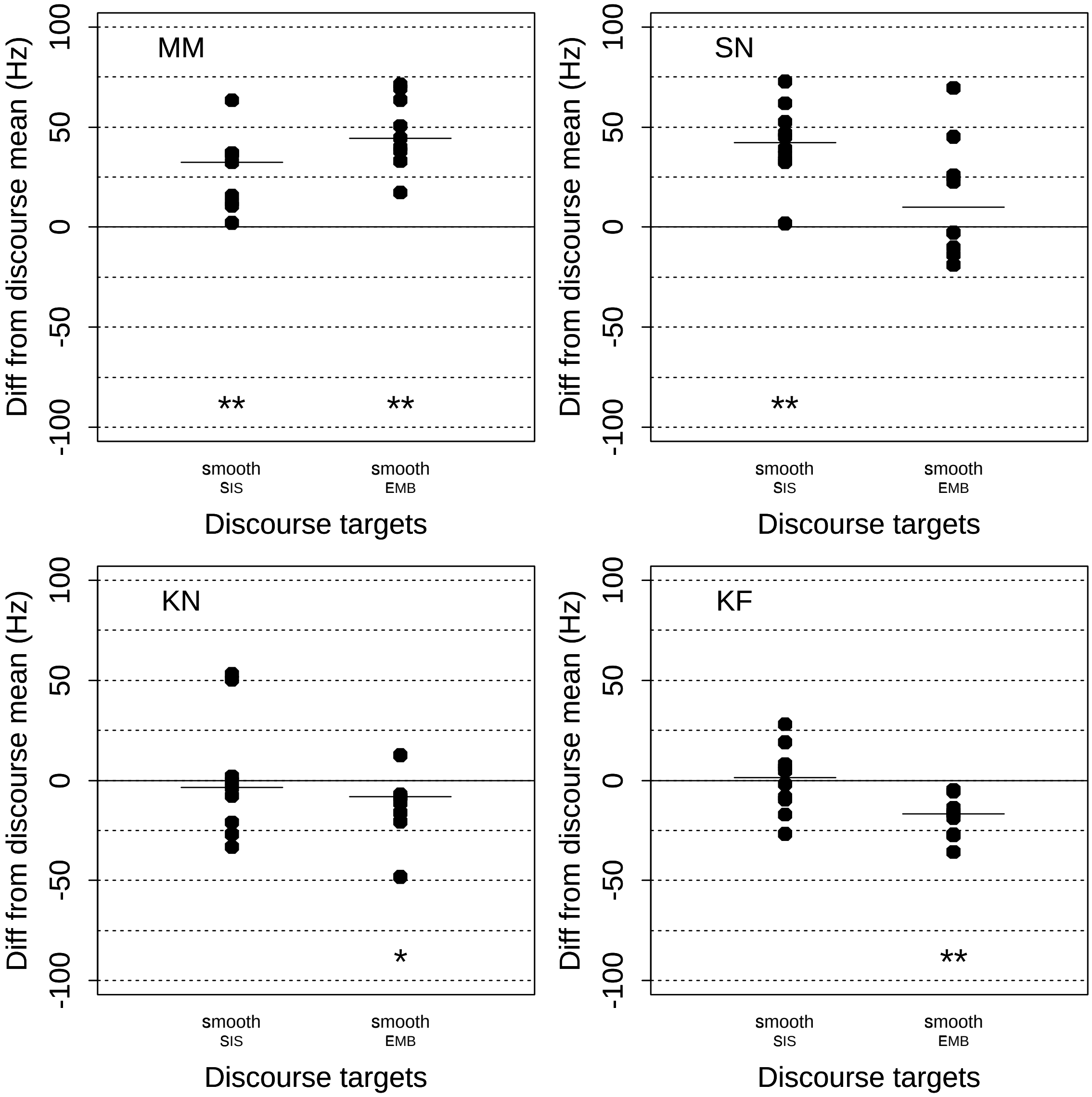

Figure 7.3: Difference measures for NP-*wa* in smooth shift transitions.

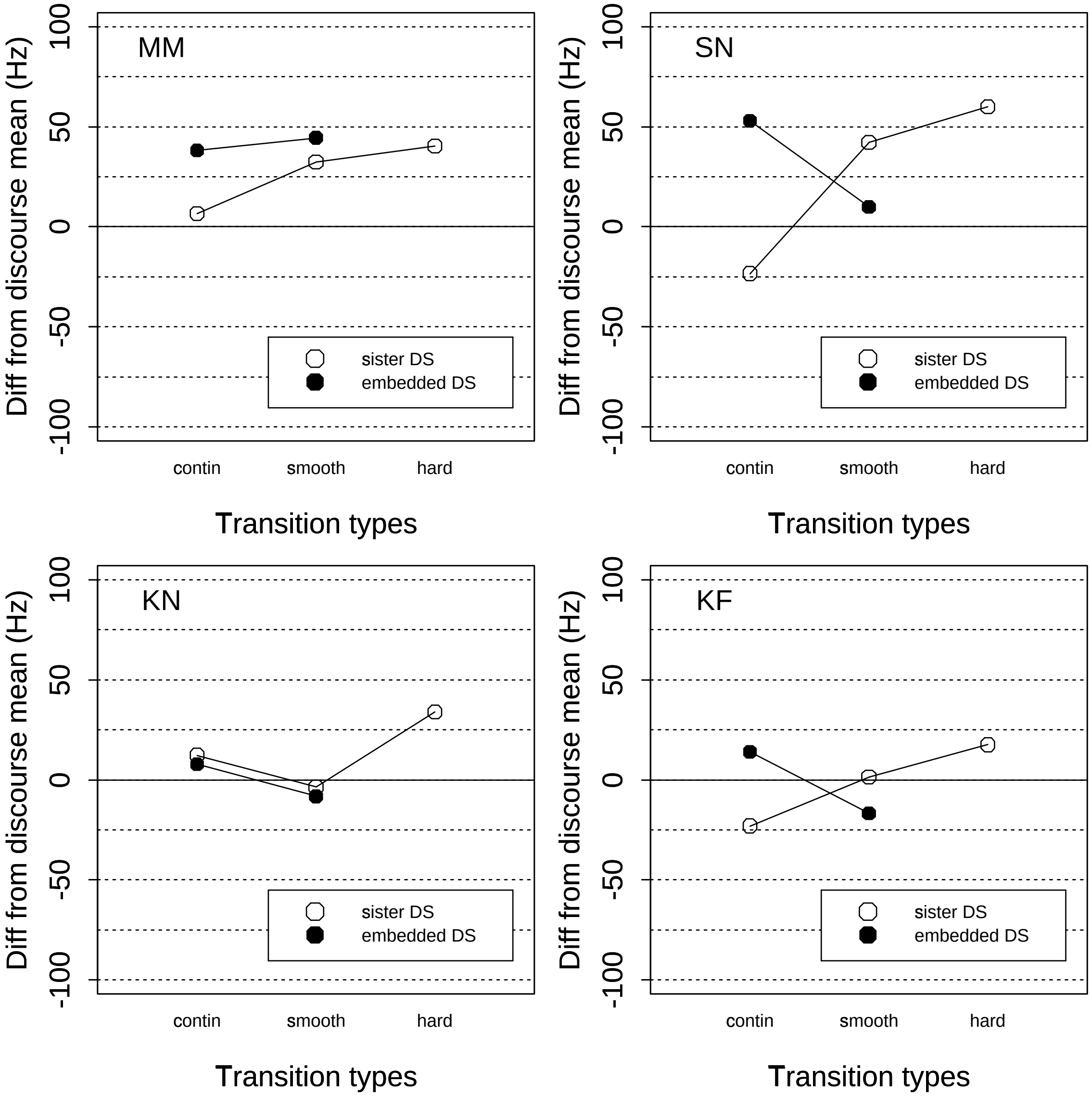

Figure 7.4: Interaction of Center transition type with hierarchical structure.

CHAPTER 8
GENERAL DISCUSSION

With large spoken language databases becoming increasingly available, we are in a unprecedented position for conducting full-scale investigations of the intonation-discourse interface. Such databases can provide the wealth of data needed to examine the role that intonation can play in cueing a variety of discourse configurations. However, linguistic analyses are impossible without a systematic tagging of a database, based on principled models of both intonational and discourse structuring. In Chapters 2 and 3, I have outlined two such models, and the review of previous studies in Chapter 4 suggests that research which merges these two approaches in the analysis of spoken language databases can yield new insights into the relation of intonation to discourse in English, Japanese, and other languages. However, the study of the intonation-discourse interface generally, and of the interface in Japanese in particular, is still in its infancy. Specifically, we have limited knowledge of how speakers can mark the structuring of their intentions over the course of a discourse, and also how they may cue the dynamic changes in the salience of entities as the discourse unfolds. The experimental study presented in this thesis is an attempt to elucidate these issues.

The previous research on Japanese outlined in Chapter 4 has contributed a great amount to our knowledge of the intonation-discourse mapping. Results from studies using a variety of intonation and discourse frameworks have enabled us to form specific predictions about how intonation may cue the intention-based structuring and attentional salience in discourse described by Grosz and colleagues. However, many of these studies have been confounded by (non-discourse) factors which we know to affect intonation in Japanese, such as downstepping, unaccented vs. accented high tone scaling, sentence-initiality effects, etc. For example, Sugito notes that the phonological downstepping of a target entity due to a preceding accented modifier confounds the comparison of 'given' vs. 'new' entities in her database [Sugi96]. Finn's comparison of NP-*wa* vs. NP-*ga* targets is confounded by numerous other factors as well, including lexical accentuation, and prosodic phrasing [Finn84]. In addition, even studies whose analyses are not confounded by these factors have left far too many unknowns. Hirose, et al. redefine the notion of discourse 'givenness' and 'importance' over the course of three studies of the same database [HSOF94, SH95, HSK96]. Venditti and Swerts examine pitch range marking on 'first-mention' vs. 'later-mention' NPs within topic units in narrative sequences, without any consideration of whether or not the NP is the local focus of attention, or of the discourse segment-internal positioning of the entity [VS96]. Even the pilot data of this thesis (see Section 4.2.5) suggest that, while NP-*wa* entities are realized in a low pitch range overall, there is variation *within* this class which may be systematic and meaningful.

The approach to database design and analysis used in this study overcomes the pitfalls of previous work, while still being feasible in terms of both time and labeling resources. It employs two well-known and widely-used models of discourse and intonational structuring, in order to provide an independent assessment of discourse structuring and global/local salience, and also to allow for appropriate comparisons of intonational configurations. The study uses semi-controlled constructed discourses in which targets are positioned in strategic locations, in order to tease apart the contributions of a number of discourse factors which are hypothesized to affect intonational realization, based in previous work in English, Japanese, and other languages. By controlling portions of the discourse database in this way, we can assure that the discourse configurations of interest will be adequately represented in the data, and also that proper intonational comparison can be made, without the threat of confounds from other factors.

Results from the database analysis reported in Chapters 6 and 7 suggest that Japanese speakers can use intonation, specifically pitch range variation, to cue intention-based discourse structuring and global and local attentional salience of discourse entities. I will briefly summarize these results below.

Japanese speakers can mark the edges of intention-based discourse segments using intonational prominence. Half of the speakers (MM and SN) marked the beginning of such units with increased pitch range, while a majority of the speakers (SN, KN and KF) marked the end of such units with lower pitch range. This pattern is consistent with previous studies of English and other languages which use intention-based and similar approaches to discourse segmentation [Leh75, Yule80, HP86, Silv87, GH92, SG94, HN96]. In addition, it is consistent with previous studies of Japanese read paragraphs and topic-unit marking in spontaneous Japanese narrative [Ven96, VS96].

In addition to marking discourse structuring, Japanese speakers also mark discourse salience by intonational means. A majority of speakers (SN, KN and KF) realize entities which are new to the global focus space using prominent intonation, in comparison with those which have been already introduced previously in the same focus space. This is consistent with the marking of the discourse-'given' vs. 'new' distinction in English, Japanese, and other languages [Hal67, Brown83, Terk84, HP86, Naka93, HSOF94, SH95, HSK96, VS96, Naka97a, Naka97b]. These studies have used a variety of definitions of discourse 'givenness', including Grosz and Sidner's global attentional salience which is examined here. However, in addition to just the given/new distinction as defined within the current focus space, this study also investigates the intonational marking of other discourse-'given' entities with respect to the operations of the focus stack. Specifically, the data show that a majority of speakers (MM, SN and KF) use prominent intonation to cue the re-introduction of entities whose antecedent has been previously popped from the focus stack. In such a case, the antecedent is not considered to be currently salient in the discourse. A comparison of the intonational realization of this discourse configuration to the one in which the antecedent has just been popped from the stack shows that a majority of speakers (MM, SN and KF) realize entities with previously-popped antecedents with a more prominent intonation relative to those with just-popped antecedents. In the latter case, the antecedent is considered to be still salient in the discourse. In addition, all speakers mark entities whose antecedent is in the current focus space (albeit before an intervening embedded segment) using non-prominent intonation, in order to cue the salience of the entity in the current discourse. These results are consistent with previous studies of English and Japanese that have reported the use of prominent intonation to mark entities which are new or re-introduced to the focus space, and non-prominent intonation to mark those whose antecedents are already in a space on the focus stack or have been just popped from the stack [DH88, VS96, Naka93, Naka97a, Naka97b].

In addition to marking global attentional salience, Japanese speakers also use intonation to cue the salience, and salience relations, of entities at a more local level. A comparison of entities in the controlled database in which the only distinguishing factor is their local salience shows that all speakers mark locally salient entities as relatively less prominent than ones which are only globally salient. That is, speakers can use intonation to cue listeners that a particular discourse entity is not only salient with respect to the current discourse, but is the entity which is the Center of attention in the current sequence of utterances. The data show that speakers can also mark the relation that this Center has to the immediately preceding context: if the Center of attention is maintained across a sequence of utterances, it is marked by intonational non-prominence. In contrast, if the local attention has been shifted from a previous Centered entity, then speakers tend to mark this using prominent intonation. This same effect has been reported in the description of pitch accent distribution on pronouns in English [Naka97a, Naka97b]. Here, the effect is observed on topic-marked NP-*wa* in Japanese discourse.

The current analysis of Centered NP-*wa* entities in Japanese suggests that there is some amount of speaker variability in the intonational marking of local salience relations. That is, speakers adopt different strategies to cue the relation that Centered entities have with respect to the preceding discourse context. For example, while all speakers mark a 'hard' shift of a local Center using prominent intonation, only two of the speakers (MM and SN) mark smooth shifts in the local Center of attention with prominence. In addition, there is a very clear interaction of Centering transition type with hierarchical discourse structure observed for two of the speakers (SN and KF). In these cases, the continuation of a Center to a sister discourse segment is marked by intonational non-prominence, while continuation to an embedded segment is marked by prominence. In contrast, a smooth shift to a sister segment is marked by prominence, while a smooth shift to an embedded segment is marked by non-prominence, which is the reverse of the pattern in the Center continuation case. This interaction is not predicted by any of the previous studies for English, Japanese, or other languages. However, this may only be because none of the previous studies have investigated Center transitions in different hierarchical configurations using a balanced experimental design such as this.

In conclusion, analysis of the constructed read speech database suggests that Japanese speakers use pitch range variation as as means to mark intonational prominence on entities in spoken language discourse. By employing the theory of discourse organization and attentional salience proposed by Grosz and colleagues, coupled with an intonational analysis free of confounding effects, this study is able to shed new light on the nature of the intonation-discourse mapping in Japanese. Since pitch accent is a lexical property of the word in Japanese, it is not available to cue discourse salience, as it does in English and other languages. However, this study shows that Japanese speakers use another means, namely pitch range variation, to cue the same discourse phenomena: to mark the dynamic changes in attentional salience of entities, at both the global and local levels, as the discourse unfolds.

APPENDIX A
CONSTRUCTED DISCOURSES

The nine Japanese discourses which were constructed for the read speech database are included in the figures below. All discourses are annotated with `WHY?` labels to indicate the intended linguistic and intentional structures. Table A.1 summarizes the identity and exact locations of the target phrases examined in this thesis.

target	identity	disc. num.	DS num.	sent. num.
init	*mana'ita-o*	5	4	13
med	*mana'ita-o*	9	4	12
fin	*mana'ita-o*	8	4	19
initDS1	*mana'ita-o*	7	1	1
initDS2	*mana'ita-o*	3	2	4
initDS3	*mana'ita-o*	1	3	7
initDS2res	*mana'ita-o*	2	2res	9
initDS5	*mana'ita-o*	4	5	20
new	*wa'in-o*	1	4	15
non-adj	*ra'amen-o*	3	4	16
non-adj(C)	*na'su-o*	4	4	18
adj	*wa'in-o*	5	2	6
adj(C)	*ma'ngo-o*	6	2	6
sameRES	*wa'rabi-o*	7	2res	9
sameRES(C)	*a'yu-o*	8	2res	13
same	*na'be-o*	2	4	13
nullINIT(a)	*ma'ngo-wa*	6	1	1
nullINIT(b)	*na'su-wa*	4	1	1
hardSIS	*ma'ngo-wa*	9	2	3
smoothSIS	*a'wabi-wa*	2	2	5
smoothEMB	*na'be-wa*	3	3	8
contSIS	*a'ji-wa*	1	2	4
contEMB	*ma'ngo-wa*	9	3	6
contRES	*a'wabi-wa*	4	2res	12
contSAME(a)	*a'ji-wa*	1	1	2
contSAME(b)	*a'wabi-wa*	4	2	5

Table A.1: Placement of targets in the Japanese discourses.

WHY? Instruct how to prepare Western steamed fish (洋風魚の煮物)

WHY? Instruct how to choose good fish

1 鯵が何といっても今の時期は旬ですから、今回はとびきりの鯵を選んで
季節の味を楽しみましょう。

2 鯵は水揚げされてからすぐの新鮮なものを使うのが理想的です。

3 とれてから1日たった魚は目玉がにごってくるので注意して選びましょう。

WHY? Instruct how to clean the fish

4 鯵は順序にそって丁寧に食べられないところを除去していきます。

5 中に卵がある場合はまずそれから取りはずします。

6 うろこ、えら、そして内臓も、完全に取りのぞきます。

WHY? Give tip on how to set chopping board for cleaning fish

7 まな板をいきなりすぐにキッチン台に置く前に下に新聞紙を敷きます。

8 そうすると落としたうろこがまわりにとびちるのを防ぐことができます。

9 これが、料理の後片付けを楽にするポイントです。

10 さて、鯵の下ごしらえにもどりましょう。

11 卵を取りはずし、うろことえらと内臓を取った後、魚の身全体を
冷たい水で洗います。

12 血や残った内臓部分も完全に洗い流してきれいになったら
下準備は終わりです。

WHY? Instruct how to assemble ingredients in the pot

13 次に必要な材料をすべて鍋に入れていきます。

14 魚を鍋に並べて水で溶いたブイヨンを身がひたひたにかくれるくらいに
注ぎ入れます。

15 ワインを少々そこに加えて、軽く鍋をゆすります。

16 このとき使うワインは、和食の煮物に使うみりんや酒やだし汁の
代わりになります。

WHY? Instruct how to cook the stew

17 さあいよいよ、鍋を火にかけましょう。

18 まずは強火で、沸騰したら中火にもどします。

19 水気がとんで魚に火がとおったら火をとめて、深めの器にもりつけます。

Figure A.1: Western-style stewed fish (discourse 1).

WHY? Instruct how to prepare special miso soup (変わり種みそ汁)

WHY? Introduce today's ingredients

1 味噌汁にも幅広いバリエーションがあります。

2 だしや味噌の種類はもちろん、具も全く自由なものを使って楽しんでみましょう。

3 あさりやしじみなど貝類の味噌汁は、だしに溶けこんだ独特の風味が魅力ですが、具えらびがついワンパターンになりがちです。

4 今日はアワビを使ってみましょう。

WHY? Instruct how to prepare abalone

5 アワビはぬめりが強く砂やごみが内部にたまりやすい構造になっていますから、下ごしらえに時間をかけてきれいにしましょう。

6 きれいに洗って、ごみを落としたら、食べられない部分をすべて取り除いて、食べやすい大きさに切ります。

WHY? Give tip on how to clean abalone

7 ところで、アワビや牡蠣をきれいに洗うコツですが、水道の蛇口の真下に貝を持って来て片手で固定し、反対側の手の親指を貝殻と身の間に差し込んでしごくようにすれば楽です。

8 ため水の中より、流水を使ったほうが、簡単にぬめりが落とせます。

9 まな板を手ごろな位置に用意して、今洗ったアワビから身の外側の食べられない部分をすべて取り除いたら、食べやすい大きさに切ります。

10 これでアワビの準備は完了です。

WHY? Instruct how to prepare the stock

11 次に鍋を用意して味噌汁のだしをとります。

12 市販のだしの素などで手を抜かずに、削り節を使います。

13 鍋を火にかけてまず水だけを沸騰させます。

14 削り節を加え、2分位したら取り出します。

WHY? Instruct how to assemble the soup

15 そして最後に、だしが用意できたら、材料を一度に加えます。

16 アワビ、わかめを鍋に入れ、一度煮立てます。

17 ここで味噌を溶かし入れ、豆腐と小口切りにした葱を加えます。

18 葱を加えた後は煮すぎて香りをとばさないように注意しましょう。

Figure A.2: Special miso soup (discourse 2).

WHY? Instruct how to prepare (Korean) chige (満腹チゲ煮込み)

WHY? Introduce the chige ingredients
1 チゲ煮込みは通常の材料以外にいろんなものを使ってバリエーションを楽しむことができます。
2 冷蔵庫に残っているありあわせの野菜と、それから今日は買い置きのラーメンを使って、ボリュームいっぱいの献立にします。
3 使う素材を決めたらまとめて、テーブルにおいておきます。

WHY? Instruct how to prepare vegetables and place in pot
4 まな板をすぐにここで用意して、野菜類の準備にかかりましょう。
5 基本的にはすべてざく切りにするだけです。
6 今日の素材はもやし、ごぼう、生しいたけ、春菊、ながねぎ、そして白菜のキムチです。
7 具だくさんになりますから、大きめの鍋を用意してください。

WHY? Give tip on what type of pot to use
8 鍋は鉄製の大きな厚手のものがよいでしょう。
9 一度熱くなったら後は調理する間中適度な温度で熱をしっかり保つからです。
10 素材にゆっくり確実に火がとおり、うまみが十分に溶け出します。

11 先ほど述べたように野菜をざく切りにします。
12 食べやすい大きさ、またはそれよりこころもち大きいくらいに切って、皿にもりあわせます。
13 野菜の準備はこれだけですべて完了です。

WHY? Instruct cook ingredients in pot
14 次にあらかじめとっておいた昆布だしを用意して火にかけ、ここに調理時間の長いものから順番に材料を加えていきます。
15 だしにまずおろしにんにくを加え、唐辛子粉を好みの分量だけ加えます。
16 ラーメンを煮立ったところで入れ、弱火にして煮ます。
17 ここでにんにくと唐辛子の風味がたちこめていれば味付けは十分です。

WHY? Instruct how to add and cook the vegetables
18 では最後に、切った野菜を一度に加え、もう一度煮立つまで待ちます。
19 野菜に火がとおってしんなりしたら食べごろです。

Figure A.3: Korean chige (discourse 3).

WHY? Instruct how to prepare Chinese stirfry (おいしい中華炒め物)

WHY? Instruct how to prepare the eggplants

1 茄子は今日紹介する炒めものに最適な素材ですが、火の通りをよくする工夫が肝心です。

2 よく洗って、5センチくらいに細長く切り、そこにさらに細かく切り目をいれます。

3 水気とアクがでるよう塩をまぶしてしばらく置いておきます。

WHY? Instruct how to prepare the abalone

4 アワビがとりわけ新鮮な季節ですからちょっとぜいたくをして生のアワビを使いましょう。

5 アワビは炒めものに適した素材ですが、下ごしらえがめんどうです。

6 ひだの部分にごみがたまりやすいですから注意しましょう。

7 水道の水を流しながらごみや砂を洗い落とします。

8 それから包丁を使って殻から身をはずします。

WHY? Give tip on how to keep knife sharp for cutting abalone

9 ここで大切なポイントですが、あらかじめ包丁はよく研いでおきましょう。

10 魚介類のうまみをのがさず素早く下ごしらえするためには包丁の切れ味が重要です。

11 普通の研ぎ石を使っても市販の研ぎ機で研いでもかまいません。

12 アワビはさっき述べたように貝類の中でも下ごしらえがとくにめんどうです。

13 完全にきれいになるまで洗ったら、身の部分を殻からはずして細かく切り刻みます。

14 そして汚れと砂が残ってないかもういちど念のため確かめながら水洗いします。

15 これでアワビの準備は完了です。

WHY? Instruct how to stirfry the ingredients

16 次に別々に下ごしらえしてあった材料を一緒に炒めます。

17 大きめの中華鍋に油を熱して、まずアワビを入れて、端がうっすら色づくくらいに炒めます。

18 茄子をすばやくそこに加えて、アワビと均一になるよう鍋の中をよくまぜます。

19 醤油、酒、ごま油を加え、いよいよ最後に、水溶きかたくり粉を流し入れます。

WHY? Instruct how to clean up after cooking shellfish

20 まな板をできるだけ早めに洗っておくのをここで忘れないようにしましょう。

21 食事が終わるまで待たずに冷水で包丁も一緒に洗ってしまいます。

22 魚介類を調理したあとの素早い後始末をおこたると、調理器具から匂いがとれなくなるので注意しましょう。

Figure A.4: Chinese stirfry (discourse 4).

```
WHY? Instruct how to prepare cheese fondue (スイスのチーズフォンデュ)
  WHY? Instruct which cheese to use
  1  エメンタールチーズはフォンデュによく使われる代表的なチーズです。
  2  他の種類のスイスチーズよりまろやかで、独特の香りや風味がワインや
     にんにくとあわせたときにとてもひきたつからです。
  3  最近では輸入食料品店でたいてい手に入ります。

  WHY? Instruct how to assemble the fondue in pot
  4  チーズフォンデュの材料はただまぜあわせるのではなく、
     順序にきちんと従って準備します。
  5  まずにんにくのかけらを鍋の内側にこすりつけて風味を移します。
  6  ワインを適度な量だけその中に加え、数分間弱火で熱して
     アルコール分をとばします。
  7  それからチーズを鍋に少量ずつ加えていきます。

     WHY? Give tip on what type of pot to use
     8  鍋はできればチーズフォンデュ専用のものを使います。
     9  内側にチーズが焦げ付かないよう加工されており、
        またコンロから食卓に移動させやすいデザインになっています。

  10 さて、にんにくとワインの準備ができたら鍋の中にチーズを
     少しずつ加えます。
  11 一度に加えると固まりができて溶けにくくなります。
  12 ひとつかみずつ、鍋の底にたまらないように注意しながら、
     慎重に溶かし入れます。

  WHY? Instruct how to prepare the bread
  13 まな板を水気のとばない位置に用意して、次にフランスパンを切ります。
  14 無理なく一口で食べられるような、手ごろな大きさに切っておきます。
  15 切ったパンをボウルに入れて、テーブルに置いておきます。

  WHY? Instruct how to serve the fondue
  16 では、フォンデュは熱いうちに、食卓に運びましょう。
  17 食べ方は簡単、各々がパンを金串にさして、
     鍋のチーズにつけて食べるだけです。
  18 他の料理の前菜としてももちろん、メインディッシュとしても
     立派な一品になります。
```

Figure A.5: Swiss cheese fondue (discourse 5).

WHY? Instruct how to prepare fruit salad（フルーツサラダ）

WHY? Instruct how to prepare the mango

1 マンゴーは独特の風味と明るい色合いがフルーツサラダにぴったりの素材です。

2 外側が柔らかく熟れて、内側がジューシーなものが最高です。

3 皮をむいて細長く切ります。

WHY? Instruct how to prepare the banana

4 次にバナナです。

5 今回は軽く外側をローストしたものを使い、香ばしさを味わいましょう。

6 マンゴーをひとまずここでボウルに移してから、バナナ２本の皮をむいて縦に２つ割りにします。

WHY? Give tip on roasting the banana

7 ところで、ローストする器具についてですが、これは特別なものでなくてもかまいません。

8 パン用のオーブントースターで十分です。

9 約５分くらいでちょうどよい具合に仕上がるでしょう。

10 ただし焼魚用のオーブングリルは匂いがつきますから避けてください。

11 さてバナナの下ごしらえにもどりましょう。

12 バナナ２本の皮をむいて縦に２つ割りにします。

13 オーブンで両側に軽く色がつく具合にローストして、冷めたら小さめのサイコロ大に切ります。

WHY? Instruct how to prepare the pineapple

14 ここで仕上げの素材にとりかかりますけど、まな板をいったんふきんでふいて、缶詰のパイナップルを他の素材と同じ様な大きさに切ります。

15 水っぽくなるのをさけるためあらかじめよく汁を切っておくとよいでしょう。

WHY? Instruct how to assemble the ingredients and serve

16 最後に、素材をすべて大きなボウルの中でまぜます。

17 マンゴー、バナナ、パイナップルともそれぞれ柔らかいですからつぶさないように気をつけてください。

18 仕上げにミントの葉があれば中央に飾ってできあがりです。

Figure A.6: Fruit salad (discourse 6).

```
WHY? Instruct how to prepare fried chicken rolls（ささ身ロールフライ）
  WHY? Instruct how to prepare the chicken
  1  まな板をまずここで準備して、鶏のささ身をうす切りにします。
  2  包丁を斜めに入れると十分な面積がとれるので後で巻くときに楽です。

  WHY? Instruct how to roll the ingredients
  3  次に巻きすを使って巻いていきます。
  4  ささ身の上にまず青じその葉をのせます。
  5  青じその上にさらにチーズをおいて、仕上げにわらびを少々のせてから
     全体を巻きます。

     WHY? Give tip on which cheese to use in roll
     6  ところで、チーズは高価な種類のものではなく、むしろ安い市販の
        スライスチーズがよく合います。
     7  これらのチーズは熱を加えたときに溶けやすいよう加工されているので
        揚げものには特に適しています。

  8  さっき述べたように、具をささ身の上にしそ、チーズの順番でまずのせます。
  9  わらびをチーズの上に少量置くようにのせたら、両手を使って
     すき間ができるだけできないように巻きます。
  10 ひと巻きしたところでいったんきつく締め、最後にもういちど締めると
     しっかり巻くことができます。

  WHY? Instruct how to coat the rolls
  11 さて、材料がきれいに巻けたら、溶き卵の中にいちどくぐらせて、
     パン粉を外側全体にまぶします。
  12 ここで中身が飛び出したりしていないか、そしてパン粉が全体をきれいに
     おおっているか確認してください。

  WHY? Instruct how to fry the rolls
  13 では最後に、油を中温に熱したところでさきほどのささ身ロールを、
     表面がキツネ色になるまで５分くらい揚げます。
  14 中までしっかり火が通ればささ身ロール全体がいっそうひきしまってきます。
```

Figure A.7: Fried chicken rolls (discourse 7).

```
WHY? Instruct how to prepare grilled fish (魚の塩焼)
  WHY? Instruct how to preheat the oven grill
  1  まず最初にオーブンの方の準備から始めましょう。
  2  底面にアルミホイルを敷いて、約２００度の温度にセットします。

  WHY? Instruct how to prepare and skewer the fish
  3  次に今日塩焼に使う鮎にとりかかります。
  4  鮎は内臓の苦味も楽しみのうちですから、あえて取りのぞきません。
  5  鮎の表面のぬめりだけを軽く水洗いします。
  6  ひれと尾に化粧塩をしますから、しっぽの部分が欠けたりしないように
     注意して洗います。
  7  用意しておいた竹串に鮎を一匹ずつ刺します。

     WHY? Give tip on how to prepare the skewers
     8  ところで、竹串はあらかじめ塩水につけておくのがポイントです。
     9  ある程度竹の表面が水分を含んでいたほうが扱うのに楽です。
     10 それに、塩のおかげで竹串の先がこげつきにくくなります。
     11 焼きもの料理の際には、このような竹串の準備を
        前もってしておく習慣をつけましょう。

  12 さて、竹串の使い方が出来上がりを左右する一番重要なところです。
  13 鮎を片方の手でしっかり押さえ、側面から反対側まで斜めに突き刺します。
  14 さらにもういちど波打つように尻尾のあたりに斜めに差し込みます。
  15 あとはひれと尾に化粧塩をまぶして、鮎の準備は完了です。

  WHY? Instruct how to grill the fish
  16 次に、串に刺した鮎をオーブンに並べます。
  17 隣同士がくっつかないように注意しましょう。
  18 約１５分、弱めの中火で焼きます。
  19 まな板を早めに洗っておきます。

  WHY? Instruct how to serve the fish
  20 さて、鮎が焼けたらいよいよ盛り付けです。
  21 鮎をオーブンからとりだして、洗っておいたまな板の上で
     丁寧に串を抜きます。
  22 串を抜いたあと魚の身がきれいな曲線を作っていれば大成功です。
  23 笹の葉をひいた皿にもりつけ、好みによりすだちを軽くしぼって召し上がれ。
```

Figure A.8: Grilled fish (discourse 8).

WHY? Instruct how to prepare Hawaiian breakfast（ハワイ風朝食メニュー）

WHY? Instruct how to prepare the macademia nuts

1 マカデミアナッツはハワイの名物で、今日の朝食メニューに
ぴったりの素材です。

2 今日は砕いたものを使いますが、始めから砕いたものが手に
入らなくても、紙袋に入れて外から固いもので叩けば簡単に砕けます。

WHY? Instruct how to prepare the mango

3 マンゴーはやはりその甘味が香ばしいナッツ類と合うので、
よく熟れたものを選んで下ごしらえを始めます。

4 マンゴーをきれいに洗って、スーパーのシールがついている場合は
跡が残らないようにていねいにはがします。

5 皮をつけたまま縦に半分に切って、種を取り除きます。

WHY? Give tip on how to remove mango pit

6 マンゴーは桃のように中央に大きな種がありますが、
取り除くのには少しコツがいります。

7 マンゴーを皮を下にしてしっかり固定し、小さめのナイフで種の回りを
こそぐようにして切ります。

8 そうしたら、種が簡単に取り除けます。

9 縦に半分にしたマンゴーは、今日のメニューでは皮をつけたまま、
他の素材をのせる器として使います。

10 切り口が変色しないようにライムを表面にしぼっておきます。

WHY? Instruct how to prepare the pineapple

11 仕上げの素材にとりかかります。

12 まな板をいったんここで洗って、パイナップルを小さく切ります。

13 もし生のハワイ産パイナップルがあればそれが一番ですが、
なければ缶詰のパイナップルでも十分です。

14 ただし水気をしっかり切ってから使うようにしましょう。

WHY? Instruct how to assemble ingredients

15 最後に、いよいよ素材をもりつけます。

16 先ほどのマンゴーの器にパイナップルを適量のせ、砕いたマカデミアナッツを
ふんだんに上からちらします。

17 好みによりライムをもっとふりかけるとさっぱりした食感が楽しめます。

Figure A.9: Hawaiian-style breakfast (discourse 9).

BIBLIOGRAPHY

[Ayers94] Gayle M. Ayers. 1994. Discourse functions of pitch range in spontaneous and read speech. *Ohio State University Working Papers in Linguistics*, 44:1–49.

[Ayers96] Gayle Ayers Elam. 1996. *Nuclear accent types and prominence: Some psycholinguistic Experiments.* PhD thesis, Ohio State University.

[BHF83] Mary E. Beckman, Susan R. Hertz, and Osamu Fujimura. 1983. SRS pitch rules for Japanese. *Working Papers of the Cornell Phonetics Laboratory*, 1:1–16.

[BP86] Mary E. Beckman and Janet B. Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook*, 3:255–309.

[BE94] Mary E. Beckman and Gayle Ayers Elam. 1994. Guidelines for ToBI labelling. Unpublished manuscript, Ohio State University. Version 3.0 March 1997. [http://ling.ohio-state.edu/Phonetics/etobi_homepage.html].

[BFOS84] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees.* Chapman and Hall.

[Bren95] Susan E. Brennan. 1995. Centering attention in discourse. *Language and Cognitive Processes*, 10(2):137–167.

[Brown83] Gillian Brown. 1983. Prosodic structure and the given/new distinction. In D. Robert Ladd and Anne Cutler, editors, *Prosody: Models and Measurements*, pages 67–78. Springer–Verlag.

[Cahn] Janet Cahn. ms. The effect of itonation on pronoun referent resolution. Unpublished manuscript.

[Cahn95] Janet Cahn. 1995. The effect of pitch accenting on pronoun referent resolution. In *Proc. of the Association for Computational Linguistics (ACL)*, pages 290–292, Cambridge, Massachusetts.

[CB96] W. Nick Campbell and Alan W. Black. 1996. CHATR: A multi-lingual speech re-sequencing synthesis system. Technical Report SP96-7, Institute of Electronics, Information and Communication Engineers (IEICE). (in Japanese).

[Chafe76] Wallace Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, and topics. In Charles N. Li, editor, *Subject and Topic*, pages 27–55. Academic Press.

[CH68] Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. MIT Press.

[CD87] Patricia M. Clancy and Pamela Downing. 1987. The use of *wa* as a cohesion marker in Japanese oral narratives. In John Hinds, Senko K. Maynard, and Shoichi Iwasaki, editors, *Perspectives on Topicalization: The Case of Japanese 'wa'*, pages 3–56. John Benjamins Publishers.

[DH88] James R. Davis and Julia Hirschberg. 1988. Assigning intonational features in synthesized spoken directions. In *Proc. of the Association for Computational Linguistics (ACL)*, pages 187–193.

[DiEug98] Barbara Di Eugenio. 1998. Centering in Italian. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 115–137. Clarendon Press.

[Finn84] Alice N. Finn. 1984. Intonational accompaniments of Japanese morphemes *wa* and *ga*. *Language and Speech*, 27(1):47–57.

[Fry00] John Fry. 2000. F0 correlates of topic and subject in spontaneous Japanese speech. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

[FS71] Hiroya Fujisaki and H. Sudo. 1971. Synthesis by rule of prosodic features of connected Japanese. In *Proc. of the International Congress on Acoustics*, pages 133–136.

[FH84] Hiroya Fujisaki and Keikichi Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan*, 5(4):233–242.

[Gib93] Jean Dickinson Gibbons. 1993. *Nonparametric Statistics*. Number 90 in Quantitative Applications in the Social Sciences. Sage Publications.

[GGG93] Peter C. Gordon, Barbara J. Grosz, and Laura A. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347.

[Grosz77] Barbara J. Grosz. 1977. The representation and use of focus in dialogue understanding. Technical report, SRI.

[GJW83] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proc. of the Association for Computational Linguistics (ACL)*, pages 44–50.

[GS86] Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

[GH92] Barbara J. Grosz and Julia Hirschberg. 1992. Some intonational characteristics of discourse structure. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 429–432, Banff, Canada.

[GJW95] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

[GS98] Barbara J. Grosz and Candace L. Sidner. 1998. Lost intuitions and forgotten intentions. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 39–51. Clarendon Press.

[Hal67] M. A. K. Halliday. 1967. Notes on transitivity and theme in English: Part 2. *Journal of Linguistics*, 3:199–244.

[Hara77] S. Haraguchi. 1977. *The Tone Pattern of Japanese: An Autosegmental Theory of Tonology.* Kaitakusha, Tokyo.

[Hat60] S. Hattori. 1960. Bun'setu to akusento (Phrasing and accent). In *Gengogaku no Hoohoo (Methods in Linguistics)*, pages 428–446. Iwanami, Tokyo. [Originally published in 1949] (in Japanese).

[Hat61] S. Hattori. 1961. Prosodeme, syllable structure and laryngeal phonemes. *Bulletin of the Summer Institute in Linguistics*, 1:1–27. International Christian University, Japan.

[HA99] Peter A. Heeman and James F. Allen. 1999. Speech repairs, intonational phrases anf discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, pages 527–571.

[HMI87] John Hinds, Senko K. Maynard, and Shoichi Iwasaki, editors. 1987. *Perspectives on Topicalization: The Case of Japanese 'wa'.* John Benjamins Publishers.

[HFK86] Keikichi Hirose, Hiroya Fujisaki, and Hisashi Kawai. 1986. Generation of prosodic symbols for rule-synthesis of connected speech of Japanese. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2415–2418, Tokyo, Japan.

[HSOF94] Keikichi Hirose, Mayumi Sakata, Masafumi Osame, and Hiroya Fujisaki. 1994. Analysis and synthesis of fundamental frequency contours for the spoken dialogue in Japanese. In *Proceedings of the ESCA Workshop on Speech Synthesis*, pages 167–170, Mohonk, New York.

[HSK96] Keikichi Hirose, Mayumi Sakata, and Hiromichi Kawanami. 1996. Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 378–381, Philadelphia, Pennsylvania.

[Hirsch93] Julia Hirschberg. 1993. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2):305–340.

[HP86] Julia Hirschberg and Janet B. Pierrehumbert. 1986. The intonational structuring of discourse. In *Proc. of the Association for Computational Linguistics (ACL)*, pages 136–144, New York, NY.

[HL87] Julia Hirschberg and Diane Litman. 1987. Now let's talk about *now*: Identifying cue phrases intonationally. In *Proc. of the Association for Computational Linguistics (ACL)*.

[HN96] Julia Hirschberg and Christine Nakatani. 1996. A prosodic analysis of discourse segments in direction–giving monologues. In *Proc. of the Association for Computational Linguistics (ACL)*, Santa Cruz, California.

[Iida98] Masayo Iida. 1998. Discourse coherence and shifting centers in Japanese texts. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 161–180. Clarendon Press.

[Jun93] Sun-Ah Jun. 1993. *The Phonetics and Phonology of Korean Prosody.* PhD thesis, Ohio State University.

[Jun] Sun-Ah Jun, editor. (forthcoming). *Prosodic Typology and Transcription: A Unified Approach.* (Collection of papers from the ICPhS 1999 satellite workshop on "Intonation: Models and ToBI Labeling". San Francisco, California).

[Kame85] Megumi Kameyama. 1985. *Zero Anaphora: The Case of Japanese.* PhD thesis, Stanford University.

[Kame86] Megumi Kameyama. 1986. A property–sharing constraint in centering. In *Proc. of the Association for Computational Linguistics (ACL)*, pages 200–206.

[Kame88] Megumi Kameyama. 1988. Japanese zero pronominal binding: Where syntax and discourse meet. In William J. Poser, editor, *Papers from the 2nd International Workshop of Japanese Syntax*, pages 47–73.

[Kawa61] Shin Kawakami. 1961. On the relationship between word-toneme and phrase-tone in Japanese language. *Onsei no Kenkyuu (Study of Sounds)*, 9:169–177.

[Kawa95] Shin Kawakami. 1995. Bunmatsu nado no jôshôchô ni tsuite (On phrase-final rising tones). In *Nihongo Akusento Ronshû (A Collection of Papers on Japanese Accent)*, pages 274–298. Kyûko Shoin, Tokyo. [Originally published in 1963] (in Japanese).

[Kori87] Shiro Kori. 1987. The tonal behavior of Osaka Japanese: An interim report. *Ohio State University Working Papers in Linguistics*, 36:31–61.

[Kuno72] Susumu Kuno. 1972. Functional sentence perspective: A case study from Japanese and English. *Linguistic Inquiry*, 111(3):269–320.

[Kuno73] Susumu Kuno. 1973. *The Structure of the Japanese Language.* MIT Press.

[KK77] Susumu Kuno and Etsuko Kaburaki. 1977. Empathy and syntax. *Linguistic Inquiry*, 8(4):627–672.

[Ladd96] D. Robert Ladd. 1996. *Intonational Phonology.* Cambridge University Press.

[Leh75] Ilse Lehiste. 1975. The phonetic structure of paragraphs. In A. Cohen and S. G. Nooteboom, editors, *Structure and Process in Speech Perception*, pages 195–203. Springer–Verlag.

[LH90] Diane Litman and Julia Hirschberg. 1990. Disambiguating cue phrases in text and speech. In *Proc. of the International Conference on Computational Linguistics (COLING)*, pages 251–256, Helsinki, Finland.

[McCaw68] James D. McCawley. 1968. *The Phonological Component of a Grammar of Japanese.* Mouton.

[Mae90] Kikuo Maekawa. 1990. Muakusento hôgen no intonêshon (Intonational characteristics of Japanese accentless dialects: A pilot study). *Onsei Gengo (Studies in Phonetics and Speech Communication)*, 4:87–110. (in Japanese).

[Naka92] Hiroshi Nakagawa. 1992. Zero pronouns as experiencer in Japanese discourse. In *Proc. of the International Conference on Computational Linguistics (COLING)*, pages 324–330, Nantes, France.

[NN94] Hiroshi Nakagawa and Shin'ichiro Nishizawa. 1994. Semantics of complex sentences in Japanese. In *Proc. of the International Conference on Computational Linguistics (COLING)*, pages 679–685, Kyoto, Japan.

[NT97] Shin'ya Nakajima and Hajime Tsukada. 1997. Prosodic features of utterances in task–oriented dialogues. In *Computing Prosody*, pages 81–93. Springer–Verlag.

[Naka93] Christine H. Nakatani. 1993. Accenting on pronouns and proper names in spontaneous narrative. In *ESCA Workshop on Prosody*, volume 41 of *Lund Working Papers*, pages 164–167, Lund, Sweden.

[Naka97a] Christine H. Nakatani. 1997. Discourse structural constraints on accent in narrative. In Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*, pages 139–156. Springer–Verlag.

[Naka97b] Christine H. Nakatani. 1997. *The computational processing of intonational prominence: A functional prosody perspective.* PhD thesis, Harvard University.

[Naka98] Christine H. Nakatani. 1998. Constituent-based accent prediction. In *Proc. of the Association for Computational Linguistics (ACL)*, pages 939–945, Montreal, Quebec.

[NGAH95] Christine H. Nakatani, Barbara J. Grosz, David D. Ahn, and Julia Hirschberg. 1995. Instructions for annotating discourses. Technical Report TR-21-95, Center for Research in Computing Technology, Harvard University.

[OC95] Yoko Ohta and Nick Campbell. 1995. Labelling of prosodic structure in Japanese. Technical Report TR-IT-0062, ATR Interpreting Telecommunications Research Laboratories.

[OPSH95] Mari Ostendorf, Patti J. Price, and Stephanie Shattuck-Hufnagel. 1995. The Boston Univeristy radio news corpus. Technical Report ECS-95-001, Boston University.

[Pass98] Rebecca J. Passonneau. 1998. Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 327–358. Clarendon Press.

[PL93] Rebecca J. Passonneau and Diane J. Litman. 1993. Intention–based segmentation: Human reliability and correlation with linguistic cues. In *Proc. of the Association for Computational Linguistics (ACL)*, pages 148–155, Columbus, Ohio.

[Pierre80] Janet B. Pierrehumbert. 1980. *The Phonetics and Phonology of English Intonation.* PhD thesis, Massachusetts Institute of Technology.

[PB88] Janet B. Pierrehumbert and Mary E. Beckman. 1988. *Japanese Tone Structure.* MIT Press.

[PH90] Janet B. Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonation contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 271–311. MIT Press.

[PBH94] John F. Pitrelli, Mary E. Beckman, and Julia Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 123–126, Yokohama, Japan.

[Poser84] William J. Poser. 1984. *The Phonetics and Phonology of Tone and Intonation in Japanese.* PhD thesis, Massachusetts Institute of Technology.

[POSHF91] Patti Price, Mari Ostendorf, Stefanie Shattuck-Hufnagel, and C. Fong. 1991. The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, 90:2956–2970.

[Prince81] Ellen F. Prince. 1981. Toward a taxonomy of given–new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press.

[Prince92] Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness, and information–status. In S. Thompson and W. Mame, editors, *Discourse Description: Diverse Analyses of a Fund Raising Text*, pages 295–325. John Benjamins Publishers.

[Ril92] Michael D. Riley. 1992. Tree–based modelling of segmental durations. In G. Bailly, C. Benoit, and T. R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pages 265–273. Elsevier Science Publishers.

[SH95] Mayumi Sakata and Keikichi Hirose. 1995. Analysis and synthesis of prosodic features in spoken dialogue of Japanese. In *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1007–1010, Madrid, Spain.

[Shima96] Kaori Shima. 1996. Anaphora resolution in Japanese: A centering approach. Unpublished Master's Thesis.

[Silv87] Kim E. A. Silverman. 1987. *The Structure and Processing of Fundamental Frequency Contours.* PhD thesis, University of Cambridge.

[SP90] Kim E. A. Silverman and Janet B. Pierrehumbert. 1990. The timing of pre-nuclear high accents in English. In John Kingston and Mary E. Beckman, editors, *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, pages 72–106. Cambridge University Press.

[SBP$^+$92] Kim E. A. Silverman, Mary Beckman, John F. Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 867–870, Banff, Canada.

[Spr98] Richard Sproat, editor. 1998. *Multilingual text–to–speech synthesis.* Kluwer Academic Publishers.

[SOH99] Richard Sproat, Mari Ostendorf, and Andrew Hunt, editors. 1999. The Need for Increased Speech Synthesis Research: Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis.

[Sugi96] Miyoko Sugito. 1996. Rôdoku ni okeru 'shinjôhô' no hyôgen (Expression of 'new information' in read speech). In *Nihonjin no eigo (The English of Japanese Speakers)*, pages 221–234. Izumi Shôin Publishers. [Originally published in 1985] (in Japanese).

[SG94] Marc Swerts and Ronald Geluykens. 1994. Prosody as a marker of information flow in spontaneous discourse. *Language and Speech*, 37(1):21–43.

[TD94] Shingo Takada and Norihisa Doi. 1994. Centering in Japanese: A step towards better interpretation of pronouns and zero-pronouns. In *Proc. of the International Conference on Computational Linguistics (COLING)*, pages 1151–1156, Kyoto, Japan.

[Terk84] Jacques Terken. 1984. The distribution of pitch accents in instructions as a function of discourse structure. *Language and Speech*, 27(3):269–289.

[TH94] Jacques Terken and Julia Hirschberg. 1994. Deaccentuation of words representing 'given' information: Effects of persistence of grammatical function and surface position. *Language and Speech*, 37(2):125–145.

[Turan98] Umit Deniz Turan. 1998. Ranking forward–looking centers in turkish: Universal and language–specific properties. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 139–160. Clarendon Press.

[Ven95] Jennifer J. Venditti. 1995. Japanese ToBI labelling guidelines. [http://ling.ohio-state.edu/Phonetics/J_ToBI/jtobi_homepage.html].

[Ven96] Jennifer J. Venditti. 1996. Effects of discourse structure on F0 in Japanese: Raising vs. lowering. Poster presented at the Annual Meeting of the Acoustical Society of America. Honolulu, Hawaii.

[Ven99] Jennifer J. Venditti. 1999. The J_ToBI model of Japanese intonation. Paper presented at the ICPhS satellite workshop on Intonation: Models and ToBI Labeling. San Francisco, California.

[VS96] Jennifer J. Venditti and Marc Swerts. 1996. Intonational cues to discourse structure in Japanese. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 725–728, Philadelphia, Pennsylvania.

[VMvS98] Jennifer J. Venditti, Kazuaki Maeda, and Jan P. H. van Santen. 1998. Modeling Japanese boundary pitch movements for speech synthesis. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pages 317–322, Jenolan Caves, Australia.

[VvS00] Jennifer J. Venditti and Jan P. H. van Santen. 2000. Superposition-based Japanese intonation synthesis using linear alignment models. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

[vS94] Jan P. H. van Santen. 1994. Assignment of segmental duration in text–to–speech synthesis. *Computer Speech and Language*, 8:95–128.

[vSM97] Jan P. H. van Santen and Bernd Möbius. 1997. A model of fundamental frequency contour alignment. In *Proceedings of the ESCA Workshop on Intonation*, pages 321–324, Athens, Greece.

[Walk98] Marilyn A. Walker. 1998. Centering, anaphora resolution, and discourse structure. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 401–435. Clarendon Press.

[WIC94] Marilyn Walker, Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232.

[WJP98a] Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors. 1998. *Centering Theory in Discourse*. Clarendon Press.

[WJP98b] Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince. 1998. Centering in naturally occurring discourse: An overview. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 1–28. Clarendon Press.

[WFH86] Anthony Woods, Paul Fletcher, and Arthur Hughes. 1986. *Statistics in Language Studies*. Cambridge University Press.

[Yule80] George Yule. 1980. Speaker's topics and major paratones. *Lingua*, 52:33–47.