

---

The Ohio State University

---

Working Papers in Linguistics No. 44

Papers from the Linguistics Laboratory

Edited by

Jennifer J. Venditti

The Ohio State University

Department of Linguistics

222 Oxley Hall  
1712 Neil Avenue  
Columbus, Ohio 43210-1298 USA  
lingadm@ling.ohio-state.edu

April 1994

- 28 \$5.00. 119 pp. (May 1983), Lawrence Clifford Schourup, *Common Discourse Particles in English Conversation*, OSU Ph.D. Dissertation.
- 29 \$5.00. 207 pp. (May 1984), edited by Arnold Zwicky and Rex Wallace: *Papers on Morphology*. Papers by Belinda Brodie, Donald Churma, Erhard Hinricks, Brian Joseph, Joel Nevis, Anne Steward, Rex Wallace, and Arnold Zwicky.
- 30 \$5.00. 203 pp. (July 1984). John A. Nerbonne, *German Temporal Semantics: Three Dimensional Tense Logic and a GPSG Fragment*, OSU Ph.D. Dissertation.
- 31 \$6.00. 194 pp. (July 1985), edited by Michael Geis: *Studies in Generalized Phrase Structure Grammar*. Papers by Belinda Brodie, Annette Bissantz, Erhard Hinricks, Michael Geis and Arnold Zwicky.
- 32 \$6.00. 162 pp. (July 1986). *Interfaces*. 14 articles by Arnold M. Zwicky concerning the interfaces between various components of grammar.
- 33 \$6.00. 159 pp. (August 1986). Joel A. Nevis, *Finnish Particle Clitics and General Clitic Theory*, OSU Ph.D. Dissertation.
- 34 \$6.00. 164 pp. (December 1986), edited by Brian Joseph: *Studies on Language Change*. Papers by Riita Blum, Mary Clark, Richard Janda, Keith Johnson, Christopher Kupec, Brian Joseph, Gina Lee, Ann Miller, Joel Nevis, and Debra Stollenwerk.
- 35 \$10.00. 214 pp. (May 1987), edited by Brian Joseph and Arnold M. Zwicky: *A Festschrift for Ilse Lehiste*. Papers by colleagues of Ilse Lehiste at The Ohio State University.
- 36 \$10.00. 140 pp. (September 1987), edited by Mary Beckman and Gina Lee: *Papers from the Linguistics Laboratory 1985-1987*. Papers by Keith Johnson, Shiro Kori, Christiane Laeufer, Gina Lee, Ann Miller, and Riita Valimaa-Blum.
- 37 \$10.00. 114 pp. (August 1989), edited by Joyce Powers, Uma Subramanian, and Arnold M. Zwicky: *Papers in Morphology and Syntax*. Papers by David Dowty, Bradley Getz, In-hee Jo, Brian Joseph, Yongkyoon No, Joyce Powers, and Arnold Zwicky.
- 38 \$10.00. 140 pp. (July 1990), edited by Gina Lee and Wayne Cowart: *Papers from the Linguistics Laboratory*. Papers by James Beale, Wayne Cowart, Kenneth deJong, Lutfi Hussein, Sun-Ah Jun, Sookhyang Lee, Brian McAdams, and Barbara Scholz.
- 39 \$15.00. 366 pp. (December 1990), edited by Brian D. Joseph and Arnold M. Zwicky: *When Verbs Collide: Papers from the 1990 Ohio State Mini-Conference on Serial Verbs* contains eighteen papers presented at the conference held at The Ohio State University, May 26-27, 1990.

- 40 \$15.00. 440 pp. (July 1992), edited by Chris Barker and David Dowty: *Proceedings of the Second Conference on Semantics and Linguistic Theory* contains twenty papers from the conference held at The Ohio State University, May 1-3, 1992.
- 41 \$12.00. 148 pp. (November 1992), edited by Elizabeth Hume: *Papers in Phonology*. Papers by Benjamin Ao, Elizabeth Hume, Nasiombe Mutonyi, David Odden, Frederick Parkinson, and Ruth Roberts.
- 42 \$15.00. 237 pp. (September 1993), edited by Andreas Kathol and Carl Pollard: *Papers in Syntax*. Papers by Christie Block, Mike Calcagno, Chan Chung, Qian Gao, Andreas Kathol, Ki-Suk Lee, Eun Jung Yoo, Jae-Hak Yoon, and a bibliography of published works in and on Head-Driven Phrase Structure Grammar.
- 43 \$12.00. 130 pp. (January 1994), edited by Sook-hyang Lee and Sun-Ah Jun: *Papers from the Linguistics Laboratory*. Papers by Benjamin Ao, Islay Cowie, Monica Crabtree, Janet Fletcher, Ken de Jong, Sun-Ah Jun, Claudia Kurz, Gina Lee, Sook-hyang Lee, Ho-hsien Pan, and Eric Vatikiotis-Bateson.
- 44 \$15.00. 223 pp. (April 1994), edited by Jennifer J. Venditti: *Papers from the Linguistics Laboratory*. Papers by Gayle M. Ayers, Mary E. Beckman, Julie E. Boland, Kim Darnell, Stefanie Jannedy, Sun-Ah Jun, Kikuo Maekawa, Mineharu Nakayama, Shu-hui Peng, and Jennifer J. Venditti.

The following issues are available through either: The National Technical Information Center, The U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22151 (PB), or ERIC document Reproduction Service (ED) Center for Applied Linguistics, 161 N. Kent St., Arlington, VA 22209.

- 2 November 1968, 128 pp. (OSU-CISRC-TR-68-3). PB-182 596.
- 3 June 1969, 181 pp. (OSU-CISRC-TR-69-4). PB-185 855.
- 4 May 1970, 164 pp. (OSU-CISRC-TR-70-26). PB-192 163.
- 6 September 1970, 132 pp. (OSU-CISRC-TR-70-12). PB-194 829.
- 7 February 1971, 243 pp. (OSU-CISRC-TR-71-7). PB-198 278.
- 8 June 1971, 197 pp. (OSU-CISRC-TR-71-7). PB-202 724.
- 9 July 1971, 232 pp. (OSU-CISRC-TR-71-8). PB-204 002.
- 11 August 1971, 167 pp. ED 062 850.
- 12 June 1972, 88 pp. (OSU-CISRC-TR-72-6). PB-210 781.
- 13 December 1972, 255 pp. ED 077 268.
- 14 April 1973, 126 pp. ED (parts only)
- 15 April 1973, 221 pp. ED 082 566.
- 16 December 1973, 119 pp. ED (parts only)

## Foreword

This volume includes papers that take an experimental approach to issues in linguistics. It is the fourth of a series of progress reports from the Linguistics Laboratory (see also OSUWPL No. 36, 38, and 43). Some of the papers are work in progress or have been presented at international conferences, while others have been submitted for publication to professional journals. I would like to thank Mary Beckman, Stefanie Jannedy, Keith Johnson, Sun-Ah Jun and Andreas Kathol for their help during the compilation of this volume. The production of this volume was supported by the National Institute on Deafness and other Communication Disorders under Grant No. 7 R29 DC01645-03 and by the Ohio State University Department of Linguistics.

Jennifer J. Venditti  
April 1994

### Phonetics and Psycholinguistics in the OSU Linguistics Laboratory

Gayle Ayers	Joyce McDonough
Mary Beckman	Chris McDougall
Julie Boland	Julie McGory
Kevin Cohen	( <i>Speech &amp; Hearing Science</i> )
Kim Damell	Jim Nemcek
Donna Erickson	Mira Oh
( <i>postdoc, Speech &amp; Hearing Science</i> )	( <i>Yeo Joo Technical College, Korea</i> )
Jamie Green	Ho-Hsien Pan
Jürgen Grün	( <i>Speech &amp; Hearing Science</i> )
Rebecca Herman	Panos Pappas
Lutfi Hussein	Shu-hui Peng
( <i>Ohio University</i> )	Roberto Perry-Carrasco
Stefanie Jannedy	( <i>Spanish &amp; Portugese</i> )
Keith Johnson	Robert Poletto
Sun-Ah Jun	R. Ruth Roberts-Kohno
( <i>currently at UCLA</i> )	Jin Shunde
Hyeon-Seok Kang	( <i>East Asian Lang. &amp; Lit.</i> )
No-Ju Kim	Liz Strand
Sook-hyang Lee	Jennifer Venditti
Ilse Lehiste	Hiroko Yamashita
Kikuo Maekawa	( <i>East Asian Lang. &amp; Lit.</i> )
( <i>visiting scholar, National Language Research Institute, Japan</i> )	

Ohio State University Working Papers in Linguistics No. 44

Papers from the Linguistics Laboratory

Table of Contents

Information concerning OSUWPL .....	iii-v
Foreword .....	vi
Phonetics and Psycholinguistics in the OSU Linguistics Laboratory .....	vi
Gayle M. Ayers     Discourse functions of pitch range in spontaneous and read speech .....	1-49
Mary E. Beckman     When is a Syllable not a Syllable? .....	50-69
Julie E. Boland     The Relationship between Syntactic and Semantic Processes in Sentence Comprehension .....	70-91
Kim Darnell, Julie Boland & Mineharu Nakayama The Influence of Orthography and Sentence Constraint on the Processing of Nouns in Japanese .....	92-104
Stefanie Jannedy     Rate Effects on German Unstressed Syllables .....	105-124
Sun-Ah Jun           Asymmetry of prosodic effects on the glottal gesture in Korean .....	125-145
Kikuo Maekawa       Is there 'dephrasing' of the accentual phrase in Japanese? .....	146-165
Shu-hui Peng         Effects of Prosodic Position and Tonal Context on Taiwanese Tones .....	166-190
Jennifer J. Venditti   The influence of syntax on prosodic structure in Japanese .....	191-223



## Discourse functions of pitch range in spontaneous and read speech\*

Gayle M. Ayers  
ayers@ling.ohio-state.edu

**Abstract:** Functions of intonation and pitch range were compared in matched spontaneous and read speech discourses. Two casual conversations were recorded, and the same speakers read scripts prepared from the original conversations. Sections with one primary speaker were examined. An intonational analysis showed that the locations of accents, phrase boundaries, and pauses differed between the spontaneous and read versions. A discourse segmentation determined that the topic structures were also different, although less so for the second conversation and its read version. Measures of pause and segment durations (as a reflection of speech rate) were made and related to the discourse segmentation units of sentence and paragraph, as well as to turn structure classifications of possible turn, 'rush through', and holding the floor. Since pitch range plays an important role in conveying the hierarchical segmentation of discourse, generally being expanded at the beginning of new topics, corresponding differences in pitch range relationships were expected. Pitch range relationships were represented in phonetic pitch trees based on phrasal peaks. These trees revealed that in addition to signaling topic structure, pitch range was also expanded for corrections and turn taking cues. In spontaneous speech, corrections and turn management disrupted pitch range cues to topic structure. However, the read versions lacked these disruptions, and the pitch range relationships reflected the topic structure more clearly. In a listening test, significantly more read utterances were misperceived as spontaneous in the conversation which had closely matching topic structures in the two versions.

### 1. Introduction

This study has two main starting points. The first is the distinction between spontaneous speech and read speech, and the second is the role of pitch range in signaling discourse structure. Spontaneous speech and read speech are generally taken to be two quite different modes of speech production and easily distinguishable from one another (Gårding, 1967; Shockey, 1974; Brown et al., 1980; Levin et al., 1982; Remez et al., 1985; Remez et al., 1986; Howell and Kadi-Hanifi, 1991; Blaauw, 1991; Blaauw, 1992). Most detailed prosodic studies of speech have been of speech read in the laboratory or newscast readings. If it is true

---

\***Acknowledgments:** Earlier versions of this work were presented at the November 1991 ASA and the January 1992 LSA conferences. The work was supported in part by a National Science Foundation Graduate Fellowship. Thanks to Jennifer Venditti for the discourse segmentation. This paper has benefited from comments made by various people, including Mary Beckman, Julie Boland, Craig Roberts, Julia Hirschberg, Bob Ladd, and participants in the phonetics seminar at the Department of Linguistics and Phonetics, Lund University, Sweden.

that spontaneous and read speech are so easily distinguishable, we can ask ourselves how much of what we learn from studying read speech holds true in spontaneous speech. However, it may be possible for prepared materials to be read

~~in a more spontaneous sounding style and not be so easily distinguishable from true~~

spontaneous speech. Such materials would have the advantage of having controlled content matter typical of read materials, but would be more like natural spontaneous speech than stereotypical read speech. This study matches spontaneous speech materials with read materials based on the spontaneous conversations and read with the aim of sounding spontaneous. I wanted to find out whether the read materials were perceived as spontaneous or read and then compare the two versions, paying particular attention to the role pitch range in signaling discourse structure in the two versions.

There are a few characteristic differences between the traditional classification of spontaneous and read speech. Read speech generally has more complex syntax than spontaneous speech because it is based on written prose. It has fewer hesitations and shorter pauses than spontaneous speech (Gårding, 1967; Brown et al., 1980). Furthermore, the distribution of pauses is different. Pauses in read speech generally align with grammatical phrases and punctuation such as periods and commas (Lehiste, 1975; Brown et al., 1980). Pauses in spontaneous speech may also lie at grammatical boundaries, but they often appear in conjunction with hesitations in the middle of syntactic constituents as the speaker searches for what to say (Gårding, 1967; Butterworth, 1975). In explicit comparisons of matched spontaneous and read speech (where the read text is based on a spontaneous discourse instead of a written text and so is not completely prototypical read speech) the read versions exhibit fewer pauses than the original spontaneous versions, and the pauses are not put in the same locations (Gårding, 1967; Shockey, 1974; Howell and Kadi-Hanifi, 1991). Howell and Kadi-Hanifi also found that readers put stresses and boundaries between tone units in different positions when reading the texts that they had produced spontaneously.

Previous studies have found that listeners are very good at correctly identifying prototypical examples of spontaneous and read speech. Levin et al. (1982) found that listeners could tell an average of 84% of the time whether an utterance was from a spontaneous story or a reading of a story. Even when the speech was low-pass filtered and none of the words were recognizable, they were identified 72% correctly. Informal classification of the differences between the two types of stories were listed as hesitations, long pauses, and non-literary words in the spontaneously told stories. Other studies have found that original spontaneous utterances and matched read productions can be distinguished even if they do not contain hesitations or lexical differences (Remez et al., 1985; Remez et al., 1986; Blaauw, 1991; Blaauw, 1992). Blaauw (1991, 1992) found that listeners could correctly identify a Dutch news reader's spontaneous answer to personal interview questions and his reading from a transcript of the interview 82% of the time when given the full sentence, and even as well as 75% of the time given just the first six syllables of a sentence. She found that the spontaneous versions had lower average F0 and less overall F0 variation, which is in direct contrast to what Remez and his colleagues found for their American English sample. These studies have found that no single acoustic aspect of the signal conveys the spontaneity reliably. If there are not simple acoustic correlates of spontaneous and read speech, perhaps a phonological analysis coupled with a pragmatic analysis will shed more light on the differences between spontaneous and read speech. It is unlikely however that such an analysis will help to explain the high accuracy that listeners have when listening to even small bits of speech or segmentally altered speech.



In spontaneous speech, the principle of turn taking is influential in the production of the talk, even if one person does most of the talking. According to Schegloff (1982:73), speech should be viewed as an interactional achievement.

The accomplishment or achievement is an interactional one. ... The production of a spate of talk by one speaker is something which involves collaboration with the other parties present, and that collaboration is interactive in character, and interlaced throughout the discourse, that is, it is an ongoing accomplishment, rather than a pact signed at the beginning, after which the discourse is produced entirely as a matter of individual effort.

Put simply, spontaneous speech is not a monologue, even if one speaker does most of the talking. A person speaks with an audience in mind, and interacts with that audience. Schegloff discusses several ways that a single speaker can end up doing most of the talking. One way is that the speaker may actively try to forestall interruption using what he calls 'rush through'. A speaker approaching a possible turn completion speeds up the pace of the talk, withholds a dropping pitch or the intake of breath, and phrases the talk to bridge what would otherwise be the juncture at the end of a unit. The speaker instead breaks in the middle of the next unit. A second way is that the speaker might be forced to continue talking because no coparticipant starts a next turn at an appropriate change of floor. This frequently shows up in the form of a slight gap of silence at the possible turn completion. False starts are also common here since the speaker takes the responsibility to continue talking even though not originally intending to. A third way is that the speaker might continue talking because the conversational partner actively passes an opportunity to produce a full turn of talk, such as by uttering a backchannel continuer (*hmm, yeah, etc.*). Spontaneous speech is interactive, and the interaction is both in speaking and listening for all participants. Clark and Schaefer (1989) describe the interactive character of speech in terms of presentations of utterances by speakers and acknowledgments of utterances by listeners. They say that 'a presentation is more than the uttering of a sentence. It is the reaction in real time of a spoken structure from which the partner can identify the words, phrases, and sentences that the contributor intended as final.' False starts, hesitations, and other disfluencies come because the speaker is preparing what to say 'on-line' and reacting to the full situation, including the other conversational participants. In summary, in spontaneous speech there is constant effort put into deciding what to say, making an opportunity to say it, and making sure that it is understood.

Read speech, however, does not involve such a complex communicative process. Prototypical read speech, which comes from a reading of a prepared written text, does not require the reader to figure out what to say next, because the text provides that. In read speech, turns are predefined by the blocks of text on the page and speakers always know when they are to speak. Blocks of text announce the turn structure, so they are not turns in the original sense. In effect (in contrast to what Schegloff says of spontaneous speech), in read speech a pact has been signed and the discourse is produced by individual effort so it is not a true interactional achievement. Even if the read speech is in the form of a multispeaker dialogue based on a spontaneous conversation, it develops more as a series of monologues instead of as a true dialogue like the original. The parts of the telling of the 'story' in the read speech seem to follow one another rather than being sensitive to the spontaneous context in which it was originally produced. The success or failure of recreating the illusion of a spontaneous conversation depends upon the quality of the acting of the readers and their ability to provide a simulation of a natural context. Read speech based on a spontaneous conversation can be seen as having a layer of complexity subtracted away from the original spontaneous conversation because the content is already prepared and the readers do not have to

create it on-line, and furthermore the readers know what order they will speak in so they do not have to negotiate for their turn.

The second large issue explored in this study is the way that discourse structure is signaled. This issue is independent of the distinction between

spontaneous and read speech. People organize what they say in terms of relationships of phrases and sentences into larger units, no matter whether they are speaking spontaneously or reading a text. This organization of phrases and sentences into larger units is called the discourse structure. Several different things have been shown to help signal discourse structure. Pauses at the end of sentences and longer pauses at the end of paragraphs have been found in read speech, dividing the speech stream up into units of various sizes and with different groupings (Lehiste, 1979; Brown, 1983; Silverman, 1987; Passenout and Litman, 1993). Speaking rate has also been shown to be related to topic units. Words at the beginning of topics are spoken more slowly and words at the end of topic units are spoken more quickly (Butterworth, 1975; Lehiste, 1980). However, this finding is contradicted by a finding that segment beginnings are faster as compared to segment endings (Grosz and Hirschberg, 1992). Amplitude also relates to topic units. Words at the beginning of topics are louder than words at the end of topics (Brown, 1983).

In addition to these temporal and amplitude cues, pitch range also plays an important role in conveying the hierarchical segmentation of discourse. This is the cue that I will be focusing on primarily, although I will also look at measures of pause durations and speech rate. Pitch range is expanded at the beginning of a new topic (Lehiste, 1975; Butterworth, 1975; Schegloff, 1979; Brazil et al., 1980; Brown, 1983) and compressed to varying degrees at the ends of phrases to reflect the degree of finality of an utterance (Hirschberg and Pierrehumbert, 1986; Silverman, 1987). Cooper and Paccia-Cooper (1980) found that boundary strengths can be reflected by height of F0 targets in the vicinity of the boundaries. Hirschberg and Pierrehumbert (1986) followed up on these observations of the way pitch range cues discourse structure in work with speech synthesis. They found that by systematically varying pitch range of phrases and pause lengths between segments they could signal various hierarchical relationships of topic and subtopic structure. Each discourse segment boundary was marked by a variation in pitch range which correlated with the segment's position in the overall discourse. Grosz and Hirschberg (1992) found in AP news stories that topic segment endings could be identified by relatively long following pauses, and segment beginnings could be identified by larger pitch range, shorter following pause, and by being louder and faster as compared to segment endings. Recall that this tempo cue was in contrast to earlier findings by Lehiste and Butterworth. My data given in Section 4 seemed to support the finding that segment beginnings are faster than segment endings. However, pitch range is not only implicated in signaling topic structure relationships. French and Local (1986) observe that pitch range is used in managing turn taking. They found the prosodic cues of interruptive turn taking to be high pitch and high intensity. Expanded pitch range also marks items as salient, things to pay attention to. Thus, it is also relevant to what Grosz and Sidner (1986) refer to as attentional structure -- what to successively pay attention to over time.

The present study compared how pitch range and intonational structure were used in matched spontaneous speech and read speech discourses. Two different two-speaker conversations were recorded, transcripts of the conversations were prepared, and the scripts were later read by the original speakers. The readers were instructed to read the scripts as if they were involved in a spontaneous conversation. Thus the read speech examined was not prototypical read speech, since it was specifically intended to be a simulation of spontaneous speech. I was interested in finding out how much of the illusion of spontaneity and interactivity

could be created in a read version of a spontaneous conversation. That is, how spontaneous could a read version of a spontaneous conversation sound. The read speech was a reorganization of the spontaneous speech (since it came from the same text), one which was free of the complexity of floor negotiations (since the turns were predefined) and false starts (since those were removed in the preparation of the text). I expected to find that the pitch range cues to topic structure in the read speech versions conformed fairly well to what has been suggested, i.e. that pitch range is expanded at the beginning of new topics and decreases for related subtopics, but that pitch range conveyed the hierarchical discourse structure of the spontaneous speech versions less well. I expected to find that real-time production phenomena such as floor negotiations, corrections, and false starts disrupt clear topic organization. These may complicate the role pitch range plays as a cue to topic structure, because they themselves may have manifestations in the pitch ranges used.

To test these hypotheses, independent discourse segmentations were made of both the spontaneous and read versions for each speaker. The segmentations were not strongly based on a specific theory of discourse, but they were based on the ideas of major topic breaks, turns, and corrections, which were given operational definitions. Pause durations were measured, and a measure was made of speaking rate. These temporal measures were compared with the discourse segmentations and related to previous findings. To see whether the difference in interactivity between a natural spontaneous text and a rehearsed read text could be captured by a symbolic prosodic analysis and an analysis of pitch range, I made a symbolic intonational analysis of the texts, which identified phrases and accented words. Some intonational indications of the interactive character of the spontaneous texts which were not present in the read texts are discussed in Section 5.3. From this prosodic analysis I took a measure of pitch range for each phrase, the peak fundamental frequency occurring on an accented word. Hierarchical pitch trees were constructed from these values, and the segmentations imposed by the pitch trees were compared with the discourse segmentation and events labeled. A perception test with the task of categorizing utterances as spontaneous or read was also carried out to see how "read" the read speech was. The results of the listening test are presented in Section 8.

## 2. Speech Material

The spontaneous speech used in this study was elicited by recording two separate casual conversations between friends. Each conversation lasted approximately 45 minutes. Both conversations were recorded in a soundproof room with a stereo microphone oriented to concentrate the two speakers' productions on different recording channels. I was a participant in each of the conversations. The first conversation was with FP, and the second conversation was with DW. All three speakers are native speakers of American English. Both FP and DW are male. Even though the conversations took place in a soundproof room, the conversations were very natural. There were no tasks to perform or restricted topic domains; the speakers just spoke about whatever they wished to talk about with each other. The conversations were as close to natural spontaneous speech as they could be, given that the participants knew they were being recorded to provide material for some sort of linguistic investigation. The sections that were chosen for analysis were late in the session and thus past any initial awkwardness or unnaturalness that may have arisen from the studio setting.

The read speech used in the investigation was based on parts of the spontaneous conversations. Approximately seven minutes of each of the original spontaneous conversations were selected to be produced as read speech. I transcribed these selections orthographically, and with the help of a colleague,

edited the transcripts to remove disfluencies such as false starts and pause filling hesitations. These editing decisions were made from the orthographic transcription alone, without direct reference to the audio recording. FP, DW, and I each punctuated our own parts in the edited transcripts. This method of editing and

assigning punctuation allowed the maximum opportunity for topic reorganization between the spontaneous and read versions since the groupings of words into phrases, sentences, and paragraphs were determined from the written word and not as a direct simulation of the spontaneous version. Allowing for the possibility of topic reorganization was important because one of the aims of the investigation was to explore the hypothesis that read versions had clearer manifestations of topic structure than the spontaneous versions. The readings were made from clean copies of the scripts which included the punctuation. We studied the scripts and read through them together before the actual recording, so the readings were well rehearsed. We tried to make the readings sound like spontaneous conversations, as if we were acting. None of the readers were trained actors. The read versions were approximately five minutes long.

Most of the decisions of what to remove from the orthographic transcription of the spontaneous speech were quite straightforward, but some of the choices made using this method did not reflect the original intentions of the speakers. Specifically, some false starts were not accurately edited. Consider the examples given in (1). Example (1a) is the orthographic transcription of part of FP's original spontaneous conversation, and example (1b) is the read production. (Pause lengths are also included in these examples, although they were not in the original orthographic transcription. They are shown in milliseconds between angle brackets.)

- (1) a. mm but I <64> had I mean the stuff he knows <583>  
is kind of amazing 'cause <1137> he does a lot of  
uh environmental impact stuff <694>
- b. but I mean the stuff he knows is kind of amazing  
because he does a lot of environmental impact  
stuff <454>

The edited orthographic transcription did not reflect that FP stopped after the word *'cause* and started over again with a new sentence *he does a lot of environmental impact stuff*. A more accurate edited and punctuated transcription would have been *The stuff he knows is kind of amazing. He does a lot of environmental impact stuff*. The punctuation in the read version made the phrase after *because* into a subordinate clause, which differs from the structure of the spontaneous production. There were also a few quiet backchannel listening or agreement noises such as *mm-hmm* and *hmm* which were not noted in the original orthographic transcription from which the read script was prepared. This omission of the listener's comments was unintentional and changed the character of the read text. Schegloff (1982:74) comments on the way omissions like these can affect a text. He says that when the behavior of the listeners are separated from the telling of a story, then the parts of the telling seem to follow each other instead of being a response to the behavior of the listeners. Thus the interactivity of the original conversation is destroyed. Since these listener responses (including eye contact, etc., in addition to backchanneling responses) are not present in a subsequent reading of the conversation, the interactive nature of the original conversation is necessarily lost to a certain extent.

The analysis concentrated on sections where FP and DW were the primary speakers in the spontaneous speech and the matching read speech. These sections

were each approximately one minute long. There were two reasons that I chose to concentrate on primarily single speaker sections. The first reason was that when a single speaker talks for a period of time, there is a chance for a topic to develop and be structured by pitch range changes. The second reason was that in sections with one primary speaker, the influence of explicit turn taking is minimized. Short turns and quick interchanges between speakers mix topic structure and turn taking. With these considerations, I expected to find the read speech to be a less complex version of the spontaneous speech, with a clearer topic structure. Then I could try to sort out the contributions of pitch range changes to topic structure from other, more interactive, functions of pitch range changes.

The texts of the conversation excerpts examined can be found in the first six figures. Figures 1 through 4 are of the spontaneous and read versions of two sections of Speaker FP's conversation, and Figures 5 and 6 are of the spontaneous and read versions of Speaker DW's conversation. In each of the figures my utterances are shown in italics and set off in shaded boxes. Silent intervals, a reflection of pauses, are shown in milliseconds between angled brackets (< >). These figures also show the discourse segmentation (see Sections 3 and 4) and intonational phrasing (see Section 5). The symbols PT, R, H, S, ,, \_\_, F, and C are the discourse segmentation codes, and the symbols |, ||, ) show the intonational phrasing.

### 3. Discourse segmentation

There are at least two levels of discourse segmentation which play a strong role in the organization of spontaneous speech. Both interactive turn taking and divisions into major and minor topics are important organizational principles of spontaneous speech. Spontaneous speech also has disruptions to the organized development of topics and turns in the form of on-line production phenomenon such as hesitations, false starts, and corrections. However, neither interactive turn taking phenomena nor hesitation phenomena such as false starts and corrections are particularly crucial to the discourse segmentation of read speech since the scripts provide the explicit turns and exactly what to say.

The principles of turn taking, topic structure, and on-line production phenomena were used as the basis for a qualitative analysis by an independent coder. This analysis then served as the reference for exploring possible acoustic correlates of each sort of phenomena. The independent coder was given an audio recording and a purely orthographic transcription of each of the texts, with no pauses or punctuation marks of any kind. She played the tape as much as she needed to label the data according to the instructions and labels described below. This was a purely auditory-perceptual analysis since she had no instrumental records of the speech. The labels were then related to acoustic measures such as pause lengths, standardized vowel durations (a reflection of speech rate), and pitch range relationships. The pause duration and speech rate results are described in Section 4, and the pitch range relationships are described in Section 7. This analysis was primarily a discourse segmentation and not a strong hierarchical discourse theory version of topic and subtopic relationships. I speculated on the subtopic structure based on the topic segmentations provided by the coder. The full text of the parts of the conversations studied are given in Figures 1 through 6 with the coder's labels. The coding scheme is described in the following paragraphs. The symbols PT, R, H, S, ,, \_\_, F, and C are the discourse segmentation codes, and the symbols |, ||, ) show the intonational phrasing (see Section 5). Silent intervals, a reflection of pauses, are also shown in milliseconds between angled brackets (< >). This study looked at the spontaneous and read versions of two different sections of Speaker FP's conversation ('College' shown in Figs. 1 and 2,

and 'Friend' shown in Figs. 3 and 4) and one section of Speaker DW's conversation ('Fernblaster' shown in Figs. 5 and 6).

For turn taking, the coder labeled possible turns (PT), rush through (R), holding the floor (H), and searching for a word (S). A possible turn was described

as the possible end of a turn, where the other participant could have started speaking. Rush through was described as a move by the speaker to speak faster and prevent the other speaker from taking a turn. Holding the floor was described as the speaker doing something to keep the floor and indicating that he had more to say. Searching for a word seemed to be a subcase of holding the floor and not reliably distinguishable from holding the floor otherwise. The results in Section 4 treat both holding the floor and searching for a word as holding the floor. The coder remarked that rush through only seemed possible between sentences, and that her percept of possible turn may have been based on the presence of a pause. She said that as a New Yorker (one who tends to trade turns rapidly and tolerate only short pauses at the change of floor) she may have put in more possible turns than the speakers themselves would have perceived, since they are from other parts of the country. Indeed she was correct, because I (one of the speakers) did not perceive as many possible turns as she did. Therefore, I have also marked where I considered the possible turns to be, which I had also done auditorily before I began the instrumental analysis. Those locations are the PTs marked with boxes around them, the ones where we both agreed that there was a possible turn change. I did not perceive any such locations in the read speech, but she did.

For topic structure, the coder labeled ends of sentences (.) and ends of paragraphs (□). Sentences and paragraphs were described loosely. Sentences could be syntactic sentence fragments as well as complete, well-formed syntactic sentences. A paragraph was described as a group of sentences that belonged together, and was possibly divided from the preceding or following paragraph by a change of topic. However, I did not try to impose any strong idea of what a change of topic might be.

For on-line production phenomena, the coder labeled false starts (F) and corrections (C). A false start was described as an incomplete sentence which was abandoned and not completed. A correction was described as a correction of a previous word or phrase -- for example, repeating a word with the correct pronunciation or using a new word or phrase after a false start. All of the corrections marked were self-corrections. The coder remarked that false starts and corrections did not really apply to the read speech data.

Generally the coder's labeling of the phenomena and mine agreed. However, there are a few points where I disagreed with her labels. My labels which disagree with hers use the same coding scheme, but the labels are circled. In FP's spontaneous version of 'College', shown in Fig. 1, I felt that there was a false start and correction between the phrases *Spanish I was uh* <661> and *necessarily had uh* <317>. In FP's spontaneous version of 'Friend' shown in Fig. 3, I strongly disagree with her labeling of the part *the stuff he knows is kind of amazing 'cause he does a lot of uh environmental impact stuff*. She marked an end of sentence after *amazing* and a rush through between *amazing* and *'cause*. I disagree that there is a sentence break there. My judgment is that the break is after *'cause*, at the long pause of 1137 ms, and that that marks the end of a false start and the beginning of a correction to the false start with the phrase *he does a lot of uh*. Otherwise our judgments were basically in agreement. She marked every instance of a repeated word as a correction, while I did not necessarily think of this kind of stuttering as a correction. We perceived hesitations and major paragraph breaks in the same places.

Speaker FP, Spontaneous: 'College'

Why was it appealing when it was on computer? <653>

uh because uh <894> || H PT

I F mean I

t- ma- }

to C make a map ||

on a computer would }

C is I

not <47> ||

nearly as much fun as <507> || H

F C to me ||

this seems very obvious <322> || [laugh]. PT

to make it on ||

F C to make it by ||

hand I

is much more fun than to make it on a computer ||, R

but anyway <553> || PT

um <523> || H PT

if you do- C if you can't see that I

then I C I don't know if I can explain it to you <634> || . PT

[laugh] so I n-)

I C knew I wasn't going to be a cartographer I

but I had no idea what I

was going to do <323> || . PT

and <45> || H

I ||

had registered for Spanish ||

simply because I had taken it for

five years in high school <461> || . PT

and <469> || H

because I was taking I

Spanish I was uh <561> || H PT

**FC** necessarily had uh <317> || H

F well H the <305> }

the advisor to f- <198> }

C fill out my schedule for the first semester said ||

why don't you take introduction <141> I

introdu- <130> }

C introductory linguistics ||

which was <53> one ninety ||, R

which is our I

two oh one <1342> ||, **PT**

and I I

took it I

with I

uh <492> || H PT

F it was taught by I

Thomas ||

Field ||, R

Dr. Thomas Field <623> || . PT

whom Mary I

knows <352> || . PT

because he I

also I

went I

to Cornell ||

F graduated from Cornell <302> || . PT

and he does stuff with um <177> || H

Occitan <50> and <855> || H PT

minority I

French languages || . PT

and speaks them well enough <109> ||

to be <35> mistaken as a native <536> || . PT

in that part of C of uh France <125> || . PT

which is amazing <692> ||, **PT**

and || H

um <918> || H

ever since I

then I knew that <170> ||

linguistics was something I was interested in <530> || . PT

and <134> || H

I never took any really hard I

core stuff there <620> || . PT

um <1078> || H

but I knew that <113> }

ling- <378> }

**FC** being a linguist is what I wanted to do <228> || . PT

I graduated from college in three years <502> || . PT

and <275> || H

almost went to graduate school except

that I realized I was con- <325> }

**FC** had no idea where I was going or <109> || H

what I was going to be doing so <383> || . R

I ended up teaching || . PT

but that <663> } H

F what I did while I was actually there is I was || H

an interdisciplinary studies major <283> || . PT

y' have any idea what that is <100> ||, **PT**

Yeah, I've heard.

Coding key:

**PT** (possible turn), **R** (rush through),

**H** (hold the floor), **S** (search for word),

. (end of sentence), **C** (end of paragraph),

**F** (false start), and **C** (correction).

Fig. 1. Discourse segmentation and coding. Speaker FP, Spontaneous: 'College'.

Speaker FP, Read: 'College'

Why wasn't it appealing when it was on computer? <340>

CI mean <313> || H PT  
to make a map on computer is not <91>  
FC nearly as much fun <184> || . PT  
to me this seems very obvious <603> || . PT  
to make it by hand is much more fun than to make it on <142>  
C on comp- <198> PT <ough 309> <72> H  
FC than to make it on computer <220> || . PT  
but anyway <444> || H PT  
if you can't |  
see that <86> || H  
then I don't know if I can explain it to you <434> || . PT  
so I knew I wasn't gonna be a cartographer <127> || PT  
but uh I had no idea what I was going to do <636> || . PT  
I had registered for Spanish |  
simply because I had taken it for five years in high school <382> || . PT  
and because I was taking Spanish || H  
the advisor |  
to fill out my schedule ||  
for the first semester |  
said <336> || H PT  
why don't you take introductory linguistics || H  
which was one ninety <88> || PT  
R which is our ||  
two oh one <538> || . PT  
and it was <195> taught by Dr. Thomas |  
Field <807> || . PT  
Thomas Field || H  
whom Mary knows |  
because he also went to Cornell <244> || H PT  
graduated from Cornell || . PT  
and he does stuff with uh |  
Occitan and minority French languages <364> || . PT  
and he speaks them well enough to be mistaken  
as a native in that part of French || PT  
which is <267> if(s)- <70> C is amazing <650> || . PT  
and ever since then <258> || H PT  
I knew linguistics was something I was interested in <158> || . PT  
and I never <334> || S  
took any really hard <230> |  
core stuff there <358> || H PT  
but I knew that being a linguist is what I wanted to do <611> || . PT  
I graduated from college |  
in three years <340> || . PT  
and almost went to graduate school || H  
except that I realized that I had no idea <inhale> || S  
where I was going ||  
or what I was going to do <559> || H . PT  
and |  
so I ended up |  
teaching <691> || . PT  
but <152> | S  
what I did while I <209> | S  
actually was there <152> || H PT  
is I was an interdisciplinary studies major || . PT  
you have any idea what that is || . PT  
Yeah, I've heard.

Fig. 2. Discourse segmentation and coding. Speaker FP, Read: 'College'.

Figures 1 and 2 show the spontaneous and read versions of the 'College' part of FP's conversation. Immediately obvious is that the divisions into paragraphs that the coder assigned are not the same for the two versions. The spontaneous version was divided into four paragraphs while the read version is divided into five paragraphs. The end of the first paragraph in the spontaneous version was also perceived as the end of a paragraph in the read version. Then there is a major departure in the paragraph divisions in the two versions. In the spontaneous version the conversation flowed without clearly perceptible breaks from one detail to the next in the second paragraph, from registering for



Speaker FP, Spontaneous: 'Friend'

a friend of mine um <216> || H  
works for NASA || R  
he's a physis- <112> |  
cphysicist <698> || H. [PT]  
and works at NASA <389> || . PT

*mm-hmm*

R or 'eCF used to work for NASA || R  
he now works for uh <963> | H PT  
S uh <98> || H  
Federal Energy Regulat- <95> |  
no <1597> || H PT  
uh S C Department of Energy <197> || . PT  
he works for Dept. of Energy <175> || . PT  
and he <248> |  
visits all the nuclear |  
power plants in the country <953> || . [PT]

*hmm*

which is |  
I suppose |  
interesting work <1250> || . [PT]

mm but I <64> |

had |

I C mean ||

the stuff he knows <583> || H PT  
is kind of amazing . R 'cause <1137> ||

(FC) he does a lot of uh | H  
environmental impact || H PT  
stuff <694> || H. [PT]  
and so <603> || H PT  
a lot of things |

that aren't |  
necessarily related to physics |  
he knows <42> || . R  
which is <98> |

FC at's <66> really interesting <1196> || . [PT]

he knows uh | H  
geography |  
and climate of |  
just about every region |  
of the United States <391> || . [PT]

*Well that's convenient if ever he wants to move  
somewhere nice when he retires or gets sick of  
nuclear energy.*

Fig. 3. Discourse segmentation and coding. Speaker FP, Spontaneous: 'Friend'.

introductory linguistics, to the teacher who taught it, to the teacher's research, to his reaction to the course. Only when he said that he graduated from college in three years did the coder say that a new paragraph had begun. In the read version of this same section, the section was divided into three different paragraphs. Essentially the points that flowed from one to the next in the spontaneous version were given stronger emphasis in the read version and were judged to be independent paragraphs. Both versions had a paragraph beginning at *I graduated from college in three years*. Another part of FP's conversation, 'Friend', is shown in Figs. 3 and 4 in both the spontaneous and read versions. Again, as in Figs. 1 and 2, the transcription codes show that the two versions had different divisions into paragraphs. In the spontaneous version there were judged to be two paragraphs, but in the read version there was only judged to be one.

Speaker FP, Read: 'Friend'


a friend of mine works for NASA || . PT  
he's a physicist ||  
and works at NASA || R  
or 'e used to work at |

FC for NASA || . PT  
he now works for the Dept. of Energy <300> || . PT  
he works for Department of Energy ||  
and he visits all the nuclear |  
power plants in the country || R  
which is ||  
I suppose ||  
interesting |  
work <453> || . PT  
but I mean ||  
the stuff he knows is kind of amazing |  
because he does a lot of |  
environmental impact stuff <454> || . PT  
and so <578> || H. PT  
a lot of things that aren't necess. related to |  
physics ||  
he knows <153> || . PT  
which is really interesting <497> || . PT  
he knows |  
geography and climate of |  
just about every region in the United States <67> || . PT

*Well, that's convenient if ever he wants to move  
somewhere nice when he retires or gets sick of  
nuclear energy.*

Fig. 4. Discourse segmentation and coding. Speaker FP, Read: 'Friend'.

Speaker DW, Spontaneous: 'Fernblaster'

  
 extending <1604> || PT  
 but I  
 I I H  
 CI <245> like to be spontaneous I  
 when I teach <677> || PT  
 uh <1628> PT  
 well I I C I I  
 kind of invent I  
 characters I  
 to help me <697> || S H, PT  
 with my teaching || H R  
 there's one <183> || PT  
*What multiple personalities? <975>*  
 it might be a manifestation of that || R  
 you never know <317> || PT  
 <inhalation> uh <548> || H  
 no I H  
 uh <754> I S H, PT  
 for <153> CF to illustrate the idea of I  
 pre-scriptive <1492> || S, PT  
 study of language I  
 uh <556> || PT  
 I always come from the standpoint of I  
 everybody's I  
 eighth I  
 grade I  
 English teacher || H R  
 Mrs. I  
 Edna Fernblaster <513> || PT  
*Fernblaster!*  
 <248> Mrs. Edna I. PT  
*She doesn't like plants or something then.*  
 <205> I I H  
 CI I don't know I PT  
 there're several || S H  
 pictures that come to mind || PT  
 some I'd <448> I S H  
 CI I'd rather not discuss in genteel company but I-  
 <laugh 420> || PT  
 uh <490> || H  
 um no uh <creak 856> || H, PT  
 no C no it's just a HC a C it's just a weird sounding name || PT  
 you know I  
 whenever I talk about somebody telling you I  
 how I  
 to speak <342> || PT  
 then I  
 you know I  
 I always I  
 come from CF well um I  
 you know Mrs. Fernblaster I  
 would tell you to do this || H, R  
 she would say never say ain't I  
 and that type of thing <1391> || PT  
 so uh <459> HPT and and CF and people <creak 316> || H  
 C people know when she's coming up I  
 anymore || PT  
 uh <60> and your eighth grade English teacher || H  
 and you hear these little titters I  
 in the back of the room <205> || PT  
 Edna Fernblaster 'n <463> || PT  
 and so I  
 you know I  
 that type of thing <125> <creak 167> || PT  
 F because <497> || H PT  
 I I- C like to use characters I  
 like that I  
 because <502> || H PT  
 it's such a basic concept I. R  
 it's nice to have something <373> || H PT  
 concrete to hang onto to help <1000> || PT  
*Right.*  
 n so <228> ||  
 there's really not been I  
 too much of a chance for <128> || S H  
 class participation as yet <642> || PT

Speaker DW, Read: 'Fernblaster'

  
 but I like to be spontaneous when I teach || PT  
 uh <395> ||  
 C I I <244> || S  
 kind of invent I  
 characters I  
 to help me with my teaching || PT  
 there's one || PT  
*What? multiple personalities? <297>*  
 it might be a manifestation of that || H R  
 uh you never know <339> || PT  
 uh no <189> || PT  
 uh to illustrate the idea of I  
 prescriptive I  
 study S of language || H  
 I always come from <256> ||  
 the standpoint of everybody's I  
 eighth grade English I  
 teacher I  
 Mrs. Edna Fernblaster || PT  
*Fernblaster!*  
 this is Edna <211> || PT  
*She doesn't like plants or something then.*  
 I don't know <263> || PT  
 there are several pictures that come to mind <91> || PT  
 eh I  
 some I'd rather not discuss in genteel company <413> || PT  
 uh <122> ||  
 no I  
 it's just a weird sounding name I  
 you know I H. R  
 whenever you talk about <389> ||  
 somebody I  
 telling you how to I  
 speak I  
 then I always say <289> ||  
 well you know Mrs. I  
 Fernblaster I  
 would tell you to do this <309> || PT  
 she would say I  
 never say ain't I  
 and that type of thing <414> || PT  
 so I  
 and <216> I S H PT  
 people know when she's coming up anymore <428> || PT  
 and your eighth grade English teacher || R  
 and you hear these little I  
 titters in the back of the room || H. R  
 Edna Fernblaster <467> || PT  
 and so you know <277> ||  
 that type of thing <103> || PT  
 I like to use characters I S  
 like that because it's I  
 such a basic concept <256> || PT  
 it's nice to have something concrete I S  
 to hang onto to help I PT  
 there's really not been too much of a chance for I  
 class participation as yet <497> || PT

Fig. 5. Discourse segmentation and coding. Speaker DW, Spontaneous: 'Fernblaster'.

Fig. 6. Discourse segmentation and coding. Speaker DW, Read: 'Fernblaster'.

The spontaneous and read versions of Speaker DW's 'Fernblaster' section shown in Figs. 5 and 6 had similar divisions into paragraphs. The first half was judged to be three paragraphs in the spontaneous version, but only one paragraph in the read version. The first two paragraph divisions in the spontaneous version align with pauses greater than 1.6 seconds. It is as if DW was taking his time making the transition from the previous part of story that he had been telling into a new aspect of the story in the spontaneous version, but that it did not take the same kind of time to make the transition in the read version. The next paragraph division after the matching halfway point was in essentially the same position, either before or after *and so you know that type of thing*, and the final paragraph division was in the same location.

#### 4. Acoustic measures of pause and speech rate

To see whether there were any easily quantifiable correlates of these "paragraphs", "possible turns", and so on, measures were made of pause durations and vowel durations (the latter as a metric of speech rate). Relatively broad phonetic transcriptions of the data were made using both auditory perception and spectrographic analysis of the speech. The spectrograms were made on a DSP Sona-Graph 5500-1, Kay Elemetrics Corporation instrument. Silent pauses and breaths were identified and their durations were measured in milliseconds, based on the spectrograms and waveforms of the speech. The pause durations are reported in the transcriptions as millisecond values shown between angle brackets (< >). Breaths are not distinguished from silent intervals, but rather are included in the pause durations given in the transcriptions. Silent intervals due to segments, such as stop closures, were not counted as pauses. I segmented the vowels guided by spectral changes between consonants and vowels. Vowel durations always included exclusively the voiced portion of a vowel, where there was an obvious voice bar. After stop consonants, the first periodic glottal pulse with both a voice bar and energy in the first formant was taken as the beginning of a vowel. Vowels were segmented from nasal and lateral contexts at the point of spectral change and damping of the first and higher formants. Voiceless vowel durations were not always possible to segment and separate from the surrounding consonants, so they were classified as voiceless and not given any duration in milliseconds.

##### 4.1 Pause measures

Previous investigators have found that read versions of spontaneous texts exhibit fewer pauses than the original versions (Gårding, 1967; Howell and Kadi-Hanifi, 1991). This was also the case for these data, as shown in Table 1. Read speech has also been found to have shorter pauses than spontaneous speech (Gårding, 1967; Butterworth, 1975), and this finding also holds for these data, also shown in Table 1. Both speakers had similar patterns of pause length distributions, with a higher mean and larger standard deviation of pause length in the spontaneous than in the read. Pause durations were significantly different between spontaneous

**Table 1**

Pause characteristics of the spontaneous and read productions by Speakers FP and DW.

	FP		DW	
	Spon	Read	Spon	Read
total number of pauses	72	46	36	23
mean duration (ms)	439	322	561	293
standard deviation (ms)	358	192	418	112

and read speech for both speakers (FP:  $t = 2.13, p < .05$ ; DW:  $t = 2.71, p < .01$ ), but the distributions of pauses within the same mode of speech was not significantly different between the speakers (spontaneous:  $t = -1.34, p > .1$ ; read:  $t = -.51, p > .1$ ).

So, pauses were longer and had more variable lengths in the spontaneous speech

than in the read speech, for both speakers. On this measure of pause length then these materials are typical of what has been found in spontaneous and read speech in the past, even though this read speech is not prototypical read speech.

Previous investigations have found pauses at the ends of sentences and longer pauses at the ends of paragraphs in read speech (Lehiste, 1979; Brown, 1983; Silverman, 1987; Grosz and Hirschberg, 1992; Passenout and Litman, 1993). Table 2 reports the pause length distributions relative to the discourse coding categories for the current data. The values in the columns of the table represent how many of the data points fall within the range of values. The first two columns are the values of pauses longer than the mean, either greater than one standard deviation above the mean (the first column) or between the mean and one standard deviation above the mean (the second column). The next two columns represent the pause durations less than the mean, either between the mean and one standard deviation below the mean (the third column) or less than one standard deviation below the mean (the fourth column). The fifth column represents those occurrences of each coding category with no following pause, and the final column represents the ones which have *um*, *uh*, or similar filled pauses, which is used mainly as an explanation for the code holding the floor, which is discussed shortly. Values in the "um" column include tokens from the first five columns, since they either did or did not have following pauses. There were not very many paragraphs in the data, but it did not seem necessary for there to be a long following pause (greater than the mean) in order for the coder to mark an end of paragraph. For Speaker FP's read version there seemed to be longer pauses at the ends of paragraphs however. Otherwise there was no compelling evidence for this claim in these data. More sentences ended with pauses than without following pauses, but again, sentences could end without a following pause.

Possible turn locations as marked by the coder correlated very closely with the presence of a following pause but had no clear correspondence with the length of the following pause. There are more pauses than turn labels, but most of her possible turn locations had a following pause. More of the possible turn locations corresponded with a following pause of longer than the mean, but she also marked possible turn locations when there was no pause at all. I marked many fewer possible turn locations than the coder did. Where I marked possible turns, the pauses were generally higher than the mean. Note, however, that as a participant in the conversation I took actual turns in the DW spontaneous conversation at places that I at later listening didn't think were appropriate as possible turn locations. That means that I took interruptive turns, and actively took the floor rather than waiting until it was given to me. Those locations had shorter than the mean pause duration, or no pause at all. I marked no possible turns at all for the read versions, because it didn't seem to me as a listener that there were possible turns in the read version.

Rush through was marked primarily on locations where there was an extremely short pause or no pause at all at the end of a sentence. This would correspond to what Schegloff (1982) called failing to pause for breath and continuing on into the next unit. Holding the floor had no clear relationship to following pause length. Sometimes there were long following pauses and sometimes no following pause at all. However, pause fillers and hesitation words like *um* and *uh* were closely linked with the label of holding the floor. There were more instances of rush through and holding the floor marked in the spontaneous speech than in the read speech, as we would expect if we view these as indications

of interaction with the other conversational participant and of the speaker having to think on-line of what to say.

**Table 2**

Numbers of each type of discourse coding for each following pause category, by speaker and mode of speech.

	very long (> 1 sd)	long (> m)	short (< m)	very short (< -1 sd)	no pause	"um"
<b>a) FP spontaneous</b>						
topic structure						
sentence end	4	9	13	1	7	0
paragraph end	1	1	3	0	1	0
turn structure						
possible turn (coder)	8	15	12	0	3	4
possible turn (author)	4	3	2	0	0	0
rush through	0	0	0	2	7	0
holding the floor	6	8	8	2	5	13
<b>b) FP read</b>						
topic structure						
sentence end	9	7	5	1	5	0
paragraph end	3	1	0	1	1	0
turn structure						
possible turn (coder)	7	11	11	3	6	0
possible turn (author)	0	0	0	0	0	0
rush through	0	0	0	1	2	0
holding the floor	1	4	5	1	4	0
<b>c) DW spontaneous</b>						
topic structure						
sentence end	5	4	8	0	7	3
paragraph end	2	0	2	0	0	1
turn structure						
possible turn (coder)	4	6	13	0	5	4
possible turn (author)	3	0	2*	0	1*	0
rush through	0	0	0	0	5	0
holding the floor	0	3	7	1	11	6
<b>d) DW read</b>						
topic structure						
sentence end	5	2	5	2	7	0
paragraph end	1	0	0	1	1	0
turn structure						
possible turn (coder)	5	2	6	2	5	0
possible turn (author)	0	0	0	0	0	0
rush through	0	0	0	0	4	0
holding the floor	0	0	1	0	4	0

\* I did not mark these as possible turns listening afterwards, but I actually took turns (of the interruptive sort) at these points.

**Table 3**  
Vowel phoneme duration means and standard deviations, by speaker and mode of speech.

a) Speaker FP

vowel phoneme	mean duration (ms)		standard deviation (ms)		number of tokens	
	Spon	Read	Spon	Read	Spon	Read
i	58.2	63.0	24.8	31.1	50	52
ɪ	52.7	54.3	32.5	28.7	140	114
u	106.8	96.5	73.7	55.5	6	4
ʊ	63.5	54.5	10.0	2.1	4	2
ɛ	75.1	78.2	41.5	43.0	31	39
ə	75.5	56.1	82.7	29.6	134	137
o	123.2	97.1	70.2	65.3	22	19
æ	117.3	97.4	69.8	42.5	41	30
a	84.2	83.1	37.6	23.4	24	21
ei	89.0	81.2	31.3	27.8	17	18
ai	115.7	88.2	38.3	31.9	30	30
au	157.5	96.7	58.7	23.0	4	3

b) Speaker DW

vowel phoneme	mean duration (ms)		standard deviation (ms)		number of tokens	
	Spon	Read	Spon	Read	Spon	Read
i	73.4	65.3	31.9	29.1	25	27
ɪ	70.3	57.2	44.7	27.3	61	54
u	66.0	69.8	24.6	25.9	5	5
ʊ	51.7	47.5	17.5	16.3	3	2
ɛ	77.5	71.4	40.5	29.3	23	26
ə	94.1	66.4	64.4	44.5	82	67
o	163.7	135.9	60.8	47.9	15	12
æ	114.0	94.7	43.5	35.9	26	19
a	90.0	106.3	34.1	38.5	10	8
ei	111.4	87.7	34.1	29.9	17	15
oi	111.0	69.0	0.0	15.6	1	2
ai	106.4	101.8	37.7	56.1	20	18
au	134.8	94.5	40.4	16.2	4	4

**4.2 Speech rate measure**

We might expect that the speech rate varies more in spontaneous speech than in read speech, as a speaker rushes to hold the floor, slows down when thinking of what to say, and the like. There have also been reports in the literature about differences in speech rate at the beginning of a paragraph and the end of a paragraph, although the reports disagree on the direction of the differences. The way I chose to look at speech rate was the durations of vowels. The faster the speech rate, the shorter the vowel duration. Looking at just the raw vowel duration can give a partial answer to the question of whether speech rate varies more in spontaneous speech. Table 3 shows the means of the measured vowel durations

and their standard deviations. Voiceless vowels, which were given a duration of 0 ms by the segmentation criteria, were not included in the values shown here. For both speakers, the majority of the vowels had larger means and larger standard deviations in the spontaneous speech than the read speech. The higher standard deviations around the vowel means in the spontaneous speech indicates that there is more overall variation in rate in the spontaneous speech than the read speech.

However, the raw vowel duration alone can only tell us so much about relative speech rate. Each vowel has an inherent duration (e.g. low vowels are longer than high vowels), and with this sort of information remaining we cannot really know if the vowel at any particular point in the discourse is long or short, unless it is compared to the average for each vowel of its type. A method for factoring out this kind of inherent phoneme duration is to convert the duration of each vowel token to a z-score (i.e. the number of standard deviation units away from the mean for that vowel phoneme -- for a full description of the method see Campbell and Isard, 1991; Campbell, 1992). In this way we can see for each particular token whether it is longer or shorter relative to the others of its class. For these data, the standardized values of each vowel were calculated separately for each speaker and each mode of speech. Segments with the mean duration for their class have z-score values of 0, longer than average segments have positive z-score values, and shorter than average segments have negative z-score values. The z-score vowel durations were used as a measure of local rate of speech. So, for example, if rush through was realized by a local increase in rate, the z-score values of the vowels in those regions would be smaller than the surrounding context. Similarly, if holding the floor was partly accomplished by extending the length of a word while the speaker thought of what else to say, we would see larger z-score values for the vowel(s) of such a word.

Table 4 shows the distribution of the z-score vowel durations for the final vowel before each of the discourse codings for each speaker and mode of speech. The long vowels (i.e. greater than the mean, with positive z-scores) are in the first two columns, and the short vowels (i.e. less than the mean, with negative z-scores) are in the next two columns. The last column is for voiceless vowels, whose durations were 0 by the segmentation criteria used. In these data, sentence initial vowels tended to be shorter than sentence final vowels, and paragraph initial vowels tend to be shorter than paragraph final vowels. This does not agree with the observations that words at the beginning of topic units are spoken more slowly than words at the end of topic units (Butterworth, 1975; Lehiste, 1980). It agrees better with the finding that topic segment beginnings were faster as compared to segment endings (Grosz and Hirschberg, 1992).

The final vowel before possible turns primarily had longer than the mean duration for vowels. So, it seems that a relatively long vowel and a following pause were good cues for the coder to decide that there was a possible turn location. Rush through seemed to have a distribution on the shorter end of the scale than possible turns. Most of the vowels before a rush through were less than one standard deviation unit above the mean (z-score greater than 1), but there were some with longer durations. So, a rush through seemed to correspond to a relatively short vowel and a short following pause. Holding the floor corresponded to long vowels; most were greater than the mean and only one token was shorter than one standard deviation unit below the mean. This seems to be a more reliable correlate of holding the floor than following pause duration, which could be very long or no pause at all.

**Table 4**  
 Numbers of each type of discourse coding for each following standardized vowel length category, by speaker and mode of speech.

	very long (> 1 sd)	long (> m)	short (< m)	very short (< -1 sd)	voiceless vowel
<b>a) FP spontaneous</b>					
topic structure					
sentence start	7	8	12	3	4
sentence end	13	6	9	1	5
paragraph start	0	2	4	0	0
paragraph end	1	3	2	0	0
turn structure					
possible turn (coder)	14	13	6	0	5
possible turn (author)	1	6	1	0	1
rush through	1	3	4	0	0
holding the floor	20	6	3	0	0
<b>b) FP read</b>					
topic structure					
sentence start	1	5	10	8	3
sentence end	16	6	2	0	3
paragraph start	0	2	1	2	1
paragraph end	3	1	1	0	1
turn structure					
possible turn (coder)	19	13	3	0	3
possible turn (author)	0	0	0	0	0
rush through	1	0	2	0	0
holding the floor	5	8	2	0	0
<b>c) DW spontaneous</b>					
topic structure					
sentence start	3	7	12	1	1
sentence end	10	2	10	1	1
paragraph start	1	4	1	0	0
paragraph end	2	0	2	1	0
turn structure					
possible turn (coder)	12	6	8	1	1
possible turn (author)	0	2*	1	1	2 <sup>#</sup>
rush through	2	0	2	1	0
holding the floor	9	6	6	1	0
<b>d) DW read</b>					
topic structure					
sentence start	6	2	9	3	1
sentence end	5	9	7	0	0
paragraph start	2	0	2	0	0
paragraph end	0	2	1	0	0
turn structure					
possible turn (coder)	6	8	6	0	0
possible turn (author)	0	0	0	0	0
rush through	1	2	1	0	0
holding the floor	2	2	1	0	0

\* both from actual turns and not perceived possible turn

# one from perceived possible turn, one from actual turn



A specific example showing the z-scores for each vowel in a matched spontaneous and read section is given in Fig. 7. The example is from Speaker FP's section 'Friend'. A partial intonational transcription of this excerpt is given in (2), where (2a) is the spontaneous and (2b) is the read version. Accented syllables are underlined, and phrases and pauses are marked in the text as previously. A complete discourse coding of the example can be found in Figs. 3 and 4. The vowel z-scores are plotted against the vowel phonemes occurring in the excerpt. The z-score values of the vowels of the spontaneous version are shown by open circles and the z-score values of the read version by filled circles. These values represent the vowel durations of the words lined up below the graph. The spontaneous and read versions are aligned with each other word-by-word and vowel-by-vowel. The x-axis tick mark labels are the spontaneous version vowels. The symbol  $\emptyset$  means that the vowel was voiceless or deleted (e.g. in *he* and *and*), and a dash shows that there was no corresponding word in the other version (e.g. there was no *at* in the spontaneous version). When a vowel was voiceless or deleted in one version relative to the other, that vowel was given a z-score of -2 to graphically show a 'very short' vowel. When a word was missing relative to the other version because of editing or reading differences, the missing vowels were given z-scores of 0 to graphically show no variation in vowel duration. These two kinds of z-score assignments were purely for display purposes and played no role in the calculations.

- (2) a. he's a physis- <112> } physicist <698> ||  
 and works at NASA <389> ||  
*mm-hmm*  
 or 'e used to work for NASA ||  
 he now works for uh <963> |  
 uh <98> || Federal Energy Regulat- <95>|  
 no <1597> || uh Department of Energy <197> ||
- b. he's a physicist ||  
 and works at NASA ||  
 or 'e used to work at } for NASA ||  
 he now works for the Department of Energy <300> ||

Fig. 7 gives the flavor of the rate variation given by this measure of relative vowel duration. The most striking differences between the two versions is where long durations, i.e. relatively slow speaking rates, were used. Very long duration vowels occurred phrase finally, for holding the floor, and for searching for a word. In the read version, the very long duration vowels were phrase final vowels, especially the two tokens of *NASA*. These are clear instances of phrase final lengthening. The phrase *he's a physicist* also shows final lengthening, although not as strikingly. The spontaneous version does not show the same clear tendency for final lengthening as the read version does. In the spontaneous version, the very long vowels were other than phrase final vowels. The final vowel in the aborted word *physic-* was very long, presumably because the speaker was trying to decide if that was what he actually wanted to say. The other two very long duration vowels in the spontaneous version (well over 2) were on the first two occurrences of the pause filler *uh*. The coder and I both marked these as holding the floor and searching for a word. This is a clear example of breaking in the middle of the next unit, both a syntactic and semantic unit, as Schegloff (1982) describes, and it shows lengthening associated with searching for upcoming words. Notice, however, that the next occurrence of holding the floor and searching, on *no uh*, that neither word had a z-score over 1, so very long durations are not necessary

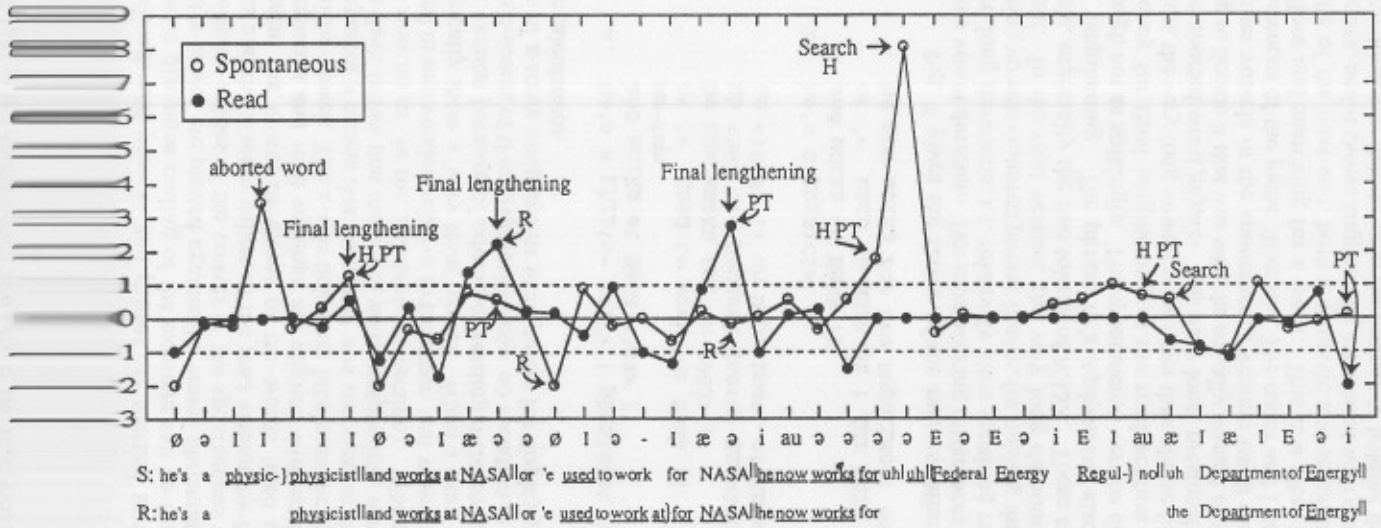


Fig. 7. Vowel duration z-scores for an excerpt from Speaker FP: 'Friend', both spontaneous and read versions. The y-axis shows the z-score value of each vowel, and the x-axis shows the vowel phonemes aligned above the appropriate text for the spontaneous (S) and read (R) versions. Underlined syllables are accented. Discourse codes are as in Figs. 1 to 6.

for holding the floor and searching to be perceived. In this specific case there was a long silence, 1597 ms, between *no* and *uh* after an incorrect mention, *Federal Energy Regulat-*, and the expectation is that he will think of the correct place and continue speaking once he has thought of it. This would be adequate in itself for the perception of holding the floor and searching for a word.

Short vowel durations correspond in some cases to the perception of rush through. In the spontaneous version the speaker uttered the phrase *or 'e used to work for NASA* with just a single accent on *used* and spoke the rest of the words relatively faster than his average rate. There was no pause for breath at the end of the phrase and the vowels were voiceless or right at or below average duration. The vowel for 'e was a voiceless vowel (hence given a -2 z-score for display purposes), and the beginning of this phrase was perceived as rushed. There was also a rush through marked at the end of this phrase ending with *NASA*, and no possible turn was judged possible there.

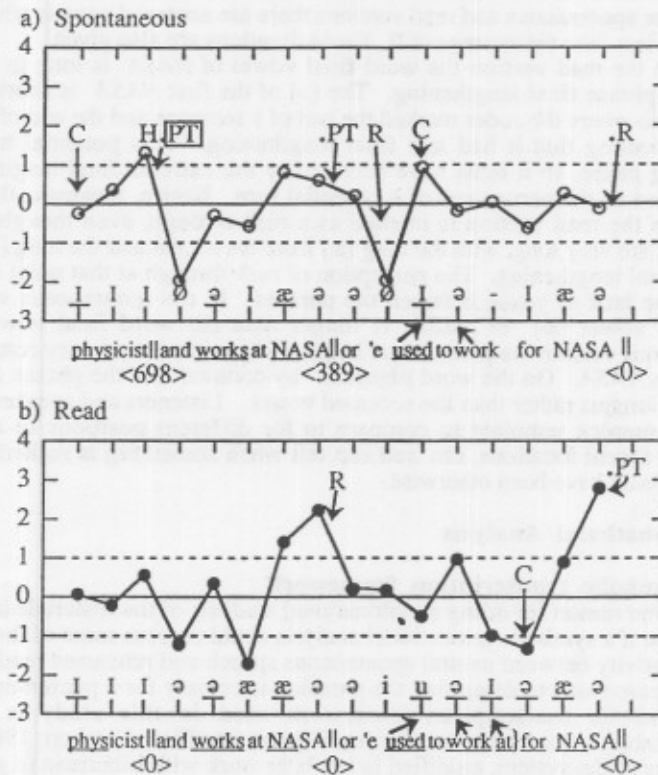
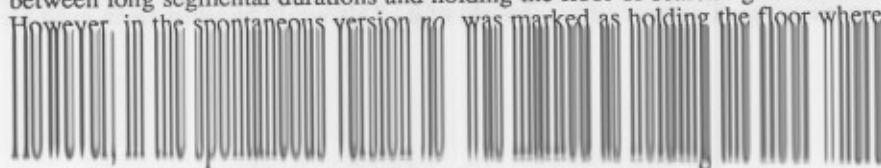


Fig. 8. Vowel duration z-scores for an excerpt from Speaker FP: 'Friend', for (a) spontaneous and (b) read versions. The y-axis shows the z-score value of each vowel, and the x-axis shows the vowel phonemes aligned above the appropriate text. Discourse codes are as in Figs. 1 to 6. Underlined vowels and syllables are accented.

In many of cases discussed above there was a correspondence between short segmental durations and the perception of rushing, and a correspondence between long segmental durations and holding the floor or searching for a word.

However, in the spontaneous version no was marked as holding the floor where



the vowel had a z-score of less than 1, and in the read version a rush through was marked at the end of the phrase *and works at NASA* where the vowel had a z-score of more than 2. Understanding things like rush through and the like from rate information are complicated by noise in the data from accents and phrase boundaries. A vowel can be long because it is an accented syllable, because it is a phrase final syllable, or because the speaker is trying to hold the floor. However, a vowel can be accented and still not be very long. Fig. 8 shows a further expansion of part of this same example. As in Fig. 7, the vowel z-scores are plotted against the vowel phonemes. In (a) the vowels durations plotted and labeled are the spontaneous vowels, and in (b) the vowels plotted and labeled are the read vowels. The underlined vowels were the vowels which were accented, and we can see that in both the spontaneous and read versions there are accented vowels which have z-scores of less than the average of 0. Pause durations are also given.

In the read version the word final vowel of *NASA* is long in both cases showing phrase final lengthening. The [ə] of the first *NASA* is shorter than the second one where the coder marked the end of a sentence and the end of a potential turn, indicating that it had less final lengthening. This potential turn had no following pause, so it must have been partly the extreme final lengthening that contributed to the perception of a potential turn. Notice, however, that the first *NASA* in the read version is marked as a rush through, even though the vowel durations are very long, with the long [æ] from the accent and the long [ə] from the phrase final lengthening. The perception of rush through at that point is probably due to the lack of pause between the phrases. In the spontaneous version, the accented vowel [æ] of *NASA* is longer than the word final vowel. In the spontaneous version the phrase final lengthening is not the primary contribution to length on *NASA*. On the word *physicist* by contrast it is the phrase final vowel which is longest rather than the accented vowel. Listeners and speakers probably have a complex template to compare to for different positions in a sentence, different accent locations, etc. and can tell when something is rushed relative to what it would have been otherwise.

## 5. Intonational Analysis

### 5.1 Symbolic transcription framework

One reason for doing an intonational analysis of the materials in this study was to see if a symbolic intonational analysis could express some of the difference in interactivity between natural spontaneous speech and rehearsed read speech. A second reason was to determine the phrasing necessary for a pitch range analysis. The symbolic transcription framework used in this study is based on Pierrehumbert's system for transcribing English (see Pierrehumbert, 1980 for some categories of the system, modified in her later work with Liberman (e.g. Liberman and Pierrehumbert, 1984) and with Beckman (e.g. Beckman and Pierrehumbert, 1986)) and the ToBI standard (Tones and Break Indices) for prosodically labeling data in American English, Australian English, and certain varieties of British English (Silverman et al., 1992). The major components of the intonational transcription system are pitch accents, phrase accents, and boundary tones. The intonational components are listed in Table 5. Only high tones (H) and low tones (L) are assumed in the phonology. Pitch accents are tones associated to certain stressed syllables. The association shows up in the time alignment of F0 to segments. There are single-tone pitch accents and bitonal accents which have two

**Table 5**

Tonal components of ToBI intonational transcription system.

Pitch accents: H\*, L\*, !H\*, L+H\* (and L+!H\*)  
(L\*+H, H+!H\* not attested in these data)

Phrase accents: L-, H-

Boundary tones: L%, H%

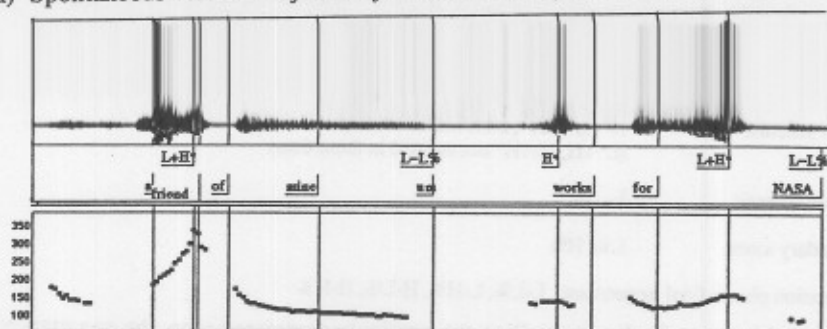
Intonation phrase final sequences: L-L%, L-H%, H-L%, H-H%

tones, with a tone leading or trailing the explicitly associated tone, the one marked with an asterisk. Words can be grouped into phrases at two levels in this system, intermediate phrases and intonational phrases. Intermediate phrases are marked by phrase accents (L- and H-), and intonational phrases by boundary tones (L% and H%). Intonational phrases can have one or more intermediate phrases. With two kinds of phrase accents and two kinds of boundary tones, there are four possible intonational phrase final sequences (L-L%, L-H%, H-L%, and H-H%). The intonational phrasing is additionally marked in the orthographic transcription of the examples to help the reader see the alignment of tones when they are of interest and to show the phrasing when the specific tones are not of interest. Intermediate phrases are marked by single vertical bars (|), intonational phrases are marked by double vertical bars (||), and intonation contours which are cut off by hesitations or restarts are marked by curly brackets ({}). The symbols for the intermediate and intonational phrase boundaries conform to the IPA guidelines for marking major and minor phrases (International Phonetics Association, 1989).

The examples in Fig. 9 illustrate some of the components of the intonational transcription system. They are examples from Speaker FP 'Friend', the spontaneous (a) and read versions (b) of the sentence *A friend of mine works for NASA*. The intonational transcription can also be found in example (3), with spontaneous (a) and read (b). The figure shows from top to bottom for each version the speech waveform, tonal transcription, word boundaries, and fundamental frequency contour. The time scale is the same for both versions, and shows that the spontaneous utterance is longer than the read version. The ends of words are marked by the labeled lines. The transcriptions are tightly linked to the fundamental frequency contour as well as to the auditory percept. H\* signifies a high target F0 on the accented syllable. The accent L+H\* is characterized by a rise from a low to a high frequency. This rise for L+H\* is seen most clearly in Fig. 9 on the word *friend* in the spontaneous version. Downstepped accents are transcribed explicitly with the downstep diacritic '!' (!H\* and L+!H\* in these data). We see downstepping in the read version of Fig. 9. The sequence of !H\* accents means that each high tone is realized on a lower pitch than it would have been were it not downstepped. There is no specific pitch movement obvious for the accented words in this example, but there is a clear percept of accent on the words *mine*, *works*, and *NASA*.

- (3) a. a friend of mine um || works for NASA ||  
           L+H\*                  L-L% H\*          L+H\* L-L%
- b. a friend of mine works for NASA ||  
           H\*                  !H\*          !H\* L-L%

a) Spontaneous version: *A friend of mine um works for NASA.*



b) Read version: *A friend of mine works for NASA.*

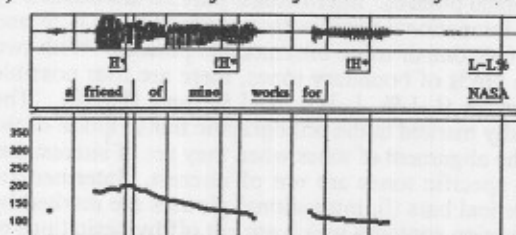


Fig. 9. Spontaneous and read versions of Speaker FP's sentence 'A friend of mine works for NASA.' Fundamental frequency in Hz.

Four types of differences between utterances can be described given this system of transcription: presence versus absence of pitch accent, type of pitch accent, phrasing, and pitch range of phrases. The examples in Fig. 9 (and example (3)) illustrate all but the first of these differences. The words *friend* and *NASA* show differences in the choice of accent type between the two versions, L+H\* in the spontaneous and H\* or !H\* in the read. There is a difference in phrasing, two intonational phrases in the spontaneous version (both ending with L-L%) in contrast to a single intonational phrase in the read version (also ending with L-L%). The two phrases in the spontaneous version gives two domains for pitch range, in contrast to the read version where there is just one. In addition, the spontaneous version was realized in a much wider pitch range in the first domain -- a peak at 337 Hz versus 200 Hz in the read version. Example (4) shows differences in presence versus absence of pitch accent, as well as phrasing and pitch range of phrases. They are again utterances from Speaker FP 'Friend', spontaneous (a) and read (b). The spontaneous version had only a single accent, on *used*, while the read version had accents on several more words. The read version has a second phrase, after the interrupted phrase correcting *at* with *for*. The pitch range of the spontaneous version, realized by the peak on *used*, was much higher than the read version -- 196 Hz as opposed to 159 Hz.

- (4) a. or 'e used to work for NASA ||  
           L+H\*  L-L%
- b. or 'e used to work at ) for NASA ||  
           L+H\*          !H\* !H\*      H\* !H\* L-L%

## 5.2 Accents

The first part of Table 6 shows the distribution of accent types in the different texts, by speaker and mode of speech. The left-hand half shows the total occurrences of each accent type, and the right-hand half shows the distribution of each accent type as a percentage of the total number of accents. The proportion of words which are accented -- that is, words which have any kind of pitch accent on them -- is similar for the spontaneous and read versions of each speaker, with a slightly smaller percentage of the words accented in the read versions. H\* was by far the most common accent type overall. There was a higher percentage of downstepped accents (!H\* and L+!H\*) in the read speech as compared to the spontaneous speech. There are other differences in the distributions, but these were the most obvious generalizations. The fact that there was a greater percentage of downstepping accents in the read version may be a cue to narrative as opposed to interactive style of communication. Bolinger (1978, p. 490) describes the downstepping intonation as "the only intonation that can be used in starting a story [of the type]: Once there was a bear. His name was Smokey." He characterizes

**Table 6**  
Accents.

a) Overall distribution of accents and accent types, by number of tokens and percentage of the total number of accents.

	number of tokens				percentage			
	FP		DW		FP		DW	
	Spon	Read	Spon	Read	Spon	Read	Spon	Read
total accented	231	203	139	120	54.4	51.8	58.5	57.7
total words	425	392	236	208				
H*	124	90	88	61	53.7	44.3	63.8	50.8
!H*	65	65	19	19	28.1	32.0	13.8	15.8
L+H*	32	34	29	30	13.9	16.7	21.0	25.0
L+!H*	5	12	2	10	2.2	5.9	1.4	8.3
L*	5	2	1	0	2.2	10.0	0.7	0.0

b) Word by word comparison of accent distributions between the spontaneous and read versions (collapsed over downstepped variation), by number of tokens and percentage.

	number of tokens		percentage	
	FP	DW	FP	DW
Words accented in both versions	152	100	54.3	64.5
a. same accents	109	68	38.9	43.9
b. different accents	43	32	15.4	20.6
Words accented in one version	128	55	45.7	35.5
a. spon unaccented	50	19	17.9	12.3
b. read unaccented	78	36	27.8	23.2
Word accent pairs	280	155		

this kind of intonational contour as being used in "self-confident" ... "narratives where a single speaker holds the floor and imposes himself on the audience". Beckman (personal communication) thinks of the downstepping contour in terms of the rather pedantic expectations set up by the act of narration where the narrator is

saying something like "This is a story. Each piece flows in a clear rhetorical succession from the last. The story structure of this discourse should give you the connections that I'm evoking by this contour ..." This difference in distribution of accent types in spontaneous and read speech could be interpreted as an essential and important difference between spontaneous and read styles of speech.

However, a numerical tally of the accent types does not reveal the whole picture of the differences in accent type distribution. The distribution of accents on individual words is also important since accent type and accent placement are pragmatic choices for highlighting or downplaying words. The second part of Table 6 shows the comparisons between spontaneous and read word pairs where at least one of the versions had a pitch accent. For Speaker FP, 54% of the words that were accented were accented in both the spontaneous and read versions, and for Speaker DW, the percentage was 65%. However, that leaves 46% and 36% of word pairs, for Speakers FP and DW respectively, where the words were only accented in one of the versions. Of those mismatched accent pairs, more of them were cases where the word was unaccented in the read version than unaccented in the spontaneous version. This goes along with the observation that a slightly smaller percentage of the words were accented in the read versions.

The choice of accent type and accent placement for particular words differed more between the two versions than the pure quantity of accents of a certain type used in a whole text. Since accent type and accent placement are pragmatic choices for highlighting or downplaying words, the two versions differ more by pragmatically determined meanings than a simple count of proportion of accent types used in a whole text can reveal. The differences in choice of accent placement between the two versions reveal differences in attentional structure, what to pay attention to over the unfolding of a discourse (Grosz and Sidner, 1986). Howell and Kadi-Hanifi (1991) also made detailed comparisons of the location of what they called "primary stress" (similar to accented words here, since they mention major pitch obstruction, loudness and length making the syllables prominent) between the spontaneous and read versions in their data and found that many of the stresses were in different positions. However, they attributed these differences to speech rate differences and made no mention of pragmatic meanings. They said that faster speech tends to have fewer stresses, and were not clear about what that might mean for differences between spontaneous and read speech. I am not aware of any other studies that make detailed comparisons between word-by-word accent locations.

### 5.3 Phrasing

The first half of Table 7 shows the number of intermediate phrases, intonational phrases, and the mean number of words and accents for each level of phrasing in the different texts, by speaker and mode of speech. The spontaneous texts have more phrases with fewer words and accents than the read texts. The read texts have fewer phrases with more words and more accents than the spontaneous texts. The number of intermediate phrases per intonational phrase is nearly identical for all of the texts. The longer phrases in the read speech may be a reflection of the fact that the words are all there and just have to be read instead of being thought through.

The second half of Table 7 shows the distribution of intonational phrase final tone sequences in the different texts. The left-hand half shows the total occurrences of each boundary tone sequence type, and the right-hand half shows the distribution of each boundary tone sequence type as a percentage of the total



number of intonational phrases. One difference in the distribution of phrase final tone sequences of spontaneous and read speech is fairly easy to interpret. Table 7 shows that while L-L% and L-H% were used heavily in both spontaneous and read speech, H-H% tended not to be used in read speech. While FP had 9 tokens of H-H% in his spontaneous speech, he had only 1 token in his read version. DW had 1 token of H-H% in his spontaneous speech and none in his read version. The contour transcribed as H-H% in this system, a phrase final high rising intonation, is a quite common American contour for inviting listener comments and indicating that the listener is to interpret what was said in terms of what follows (the situational context or the following utterance), and is the standard intonation for a yes/no question. Sacks and Schegloff (1979) calls it a 'try marker'; Clark and Schaefer (1989) calls it a 'trial constituent' when presenting a name or description that the speaker is not sure is factually correct or entirely comprehensible, and Pierrehumbert and Hirschberg (1990) calls it 'forward looking' and 'interpreting in respect to what follows'. The occurrences of H-H% in the spontaneous versions seem to be one reflection of the interaction between speaker and hearer when the speaker is producing an utterance with the hearer in mind. Note that these H-H% all occurred within the sections with one primary speaker which I was examining.

FP's single H-H% in the read speech was an explicit question, and yes/no questions typically have that pattern in American English. The intonation of the question was realized phonologically identically in the two versions. The example is given in (5). (5a) is the spontaneous version, and (5b) is the read version. The

**Table 7**  
Boundary types.

a) Phrasing statistics.

	FP		DW	
	Spon	Read	Spon	Read
intermediate phrases	121	78	69	54
mean words/phrase	3.5	5.0	3.4	3.9
mean accents/phrase	1.9	2.6	2.0	2.2
intonational phrases	73	54	48	37
mean words/phrase	5.8	7.3	4.9	5.6
mean accents/phrase	3.2	3.8	2.9	3.2
mean intermediate phrases per intonational phrase	1.7	1.4	1.4	1.5

b) Overall distribution of intonation boundary tone sequences, by number of tokens and percentage.

	number of tokens				percentage			
	FP		DW		FP		DW	
	Spon	Read	Spon	Read	Spon	Read	Spon	Read
L-L%	36	26	33	30	49.3	48.1	68.8	81.1
L-H%	27	26	11	7	37.0	48.1	22.9	18.9
H-L%	1	1	3	0	1.4	1.9	6.3	0.0
H-H%	9	1	1	0	12.3	1.9	2.1	0.0
total	73	54	48	37				

high rising tone sequence  $H^* H-H\%$  is underlined for ease of comparison. My utterances are shown in italics with sentence punctuation.

(5) a. what I did while I was actually there is I was ||

an interdisciplinary studies major <283> ||

$H^*$   $H^*$   $H^* H-H\%$   
you have any idea what that is <100> ||

$H^*$   $H^*$   $H^* H-H\%$   
*Yeah, I've heard.*

b. but <152> | what I did while I <209> |

actually was there <152> ||  
is I was an interdisciplinary studies major ||

$L+H^*$   $!H^*$   $!H^* L-H\%$   
you have any idea what that is ||

$H^*$   $H^*$   $H^* H-H\%$   
*Yeah, I've heard.*

Example (5) also shows another occurrence of  $H-H\%$  in the spontaneous version, in the phrase before the explicit question. This  $H-H\%$  is a reflection of the interaction between speaker and hearer. It is asking a question already, prefiguring the explicit question to come. However, in the read version, the first part is presented as a statement and only the explicit question has final rising intonation. A similar thing happens in example (6), one of DW's spontaneous utterances. The final high rise on *teaching* is as if to say, 'do you follow what I'm saying?', 'do you understand how inventing characters can help with teaching?'. Speaker FP also used the  $H-H\%$  in example (7) as an indication that he wondered whether he remembered correctly that his friend is a physicist. Examples (8) and (9) seem to be instances of FP using  $H-H\%$  to invite me to comment on what he has said or make some sort of response. Both the coder and I marked possible turns after the  $H-H\%$  in examples (8) and (9). All of these examples seem to be implicit questions

(6) I | kind of invent || characters ||  
to help me <697> || with my teaching ||  
 $H^*$   $!H^* H-H\%$

(7) he's a physis- <112> } physicist <698> ||  
 $H^*$   $H$   $L-$   $H^* H-H\%$   
and works at NASA <389> ||  
 $H^*$   $H^* L-L\%$

*Mm-hmm.*

or 'e used to work for NASA ||  
 $L+H^*$   $L-L\%$

(8) which is | I suppose | interesting work <1250> ||  
 $H^*$   $L-$   $L^*$   $L-H^*$   $H^* H-H\%$   
mm but I <64> } had | I mean ||  
the stuff he knows <583> ||  
is kind of amazing 'cause <1137> ||

(9) he knows uh | geography | and climate of |  
just about every region | of the United States <391> ||  
 $H^*$   $H^* H-H\%$

*Well that's convenient if ever he wants to move  
somewhere nice when he retires or gets sick of  
nuclear energy.*

such as 'do you understand what I'm saying?', 'did I say that right?', etc.

All of these examples were realized as L-L% or L-H% in the read productions, except for the explicit question of example (5). These reflections of the dialogue structure of the original conversation were eliminated in the read version when H-H% was replaced by L-L% or L-H%, making the read version more like coordinated monologues rather than a true dialogue. There was no grounding or checking to see that the listener understood by the use of the H-H% high rising contour. Thus there was interaction with the listener in the spontaneous versions which was missing in the read versions. The symbolic intonation transcription captured this reflection of the difference in interactivity of the spontaneous and read texts.

## 6. Pitch Range

### 6.1 Peak fundamental frequency as an acoustic measure of pitch range

Sections 6 and 7 specifically address the role of pitch range in structuring discourse. This section describes the measure of pitch range used, and Section 7 discusses how this measure relates to the previously determined discourse structures for the spontaneous and read texts. Evidence from downstep, prominence relations, and asides show that the intermediate phrase is an appropriate domain for local pitch range (Lieberman and Pierrehumbert, 1984; Beckman and Pierrehumbert, 1986; Grosz and Hirschberg, 1992; Silverman et al., 1992). Therefore, based on the intonational analysis of each text, I took the peak F0 of each intermediate phrase as an acoustic measure of local pitch range. The peak was taken to lie on an accented word and not on a phrase tone (H- phrase accent or H% boundary tone) if that happened to be the highest point in the pitch contour. Phrase tones were excluded as a measure of the pitch range because the phonological upstep of boundary tones after a H- would artifactually inflate the pitch range estimate by the amount of the upstep. The peaks on *friend* and *NASA* in Fig. 9, spontaneous version, are examples of such peaks. To minimize the effects of segmental perturbation and to provide a consistent measurement criteria, I measured the frequency at the point in time when the vowel of the accented syllable is at its maximum intensity (Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992). This measure treats local pitch as a continuous variable, allowing any value of F0 and not just a discrete number of pitch levels.

Fig. 10 plots the F0 peaks of all intermediate phrases for each speaker, in a frequency histogram. The spontaneous tokens are in dark gray. The mean F0 peak is significantly higher for Speaker FP in the read version (spontaneous mean: 131.6 Hz, read mean: 144.3 Hz;  $t=-2.97$ ,  $p<.01$ ), but there is no significant difference in the distribution of peak F0 for Speaker DW (spontaneous mean: 120.8 Hz, read mean: 115.7 Hz;  $t=1.01$ ,  $p>.1$ ). The pure distribution of peak frequency alone then cannot be a reliable characteristic of the difference between spontaneous and read speech because there was no consistent difference in the means. For Speaker FP the read version had a significantly higher mean, although there was an extensive overlap in distribution, and for Speaker DW the spontaneous version had a higher, but not significantly different, mean. This is in keeping with earlier results for average frequency and frequency variation (or range) (Remez et al., 1985; Remez et al., 1986; Blaauw, 1991; Blaauw, 1992). This measure looks at F0 in a global way and neglects the potential organizational principles of pitch range changes over time. Only by looking at the F0 peaks over time can we hope to see how pitch range may be used to signal discourse organization, and perhaps discover differences between uses of pitch range changes in spontaneous and read speech.

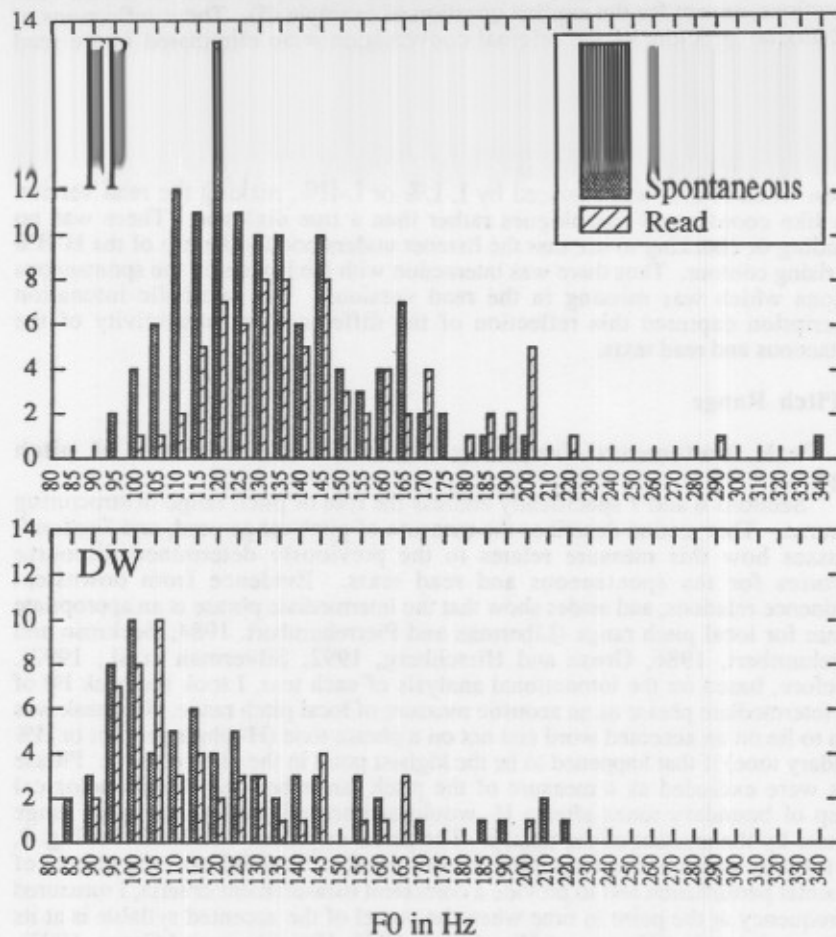


Fig. 10. Histograms of peak F0 of each intermediate phrase for Speakers FP and DW, spontaneous and read versions.

## 6.2 Local prominence

One thing increasing pitch range is used for is to increase the salience of a phrase or word in a phrase (Pierrehumbert, 1980). In these data of two male speakers, it seemed that accents which have peak F0 values of 150 Hz or more were especially salient. The value 150 Hz was one that I chose based on my impressions; subjectively those words seemed especially salient because of the high pitch. I considered these accents to be realized in an 'expanded' pitch range. No systematic perceptual testing was done in this study, but Ladd (1993 in press) in some preliminary experiments finds a difference in perception of high tones beginning at approximately 150 Hz as well for male speakers. Even though I have been viewing F0 peaks as lying on a continuous scale, it is possible that there are categorical aspects to the distribution as well, such as Ladd's overhigh tone, or uses of 'expanded' pitch range.

Fig. 11 shows an abstract representation of the expanded pitch range in a short discourse segment taken from the FP conversation. The boldface underlined words are the peak accents (i.e. the accents realized with the highest F0 in an

Speaker FP, Spontaneous:

a **friend** of mine um <216> || works for **NASA** ||.  
he's a physis- <112> | **physicist** <698> ||,  
and works at NASA <389> ||.  
[mm-hmm]  
or 'e **used** to work for NASA ||.  
he now works for uh <963> | uh <98> || Federal Energy Regulat- <95> |  
no <1597> || uh **Department** of Energy <197> ||.  
he works for Department of Energy <175> ||.  
and he <248> | visits all the nuclear | power plants in the country <953> ||.  
[hmm]  
which is | I suppose | interesting **work** <1250> ||.

Speaker FP, Read:

a **friend** of mine works for NASA ||.  
he's a **physicist** || and works at NASA || or 'e **used** to work at | for NASA ||.  
he **now** works for the Department of Energy <300> ||.  
he **works** for Department of Energy || and he visits all the nuclear | power  
plants in the country || which is || I suppose || interesting | work <453> ||.

Fig. 11. Expanded range for matched spontaneous and read excerpts of Speaker FP's conversation 'Friend'. Boldface underlined words were realized with F0 peaks of 150 Hz or greater.

intermediate phrase) with F0 of more than 150 Hz. These high pitches draw attention to the words or phrases, perhaps as a concrete reflection of something corresponding to Grosz and Sidner's attentional structure (Grosz and Sidner, 1986). The words in expanded pitch range (that is, the boldface underlined ones) were not the same ones in the two versions. If we consider pitch range as one reflection of focus in the local attentional space, the two versions had a different pragmatic or attentional structure, since the spontaneous version focused on place and the read version on time. Thus, even though the sentences in the two versions matched lexically and syntactically, the points that were made salient over the unfolding of the discourse differed. That is true even if it is not words alone but phrases which are made prominent. However, pitch range is implicated in more than just local prominence. It also participates in topic organization and turn structure. I examine these influences in the texts with the help of the hierarchical pitch tree explained in the next section.

### 6.3 The hierarchical pitch tree

The observations of pitch range and discourse hierarchy and turn taking cues discussed in the introduction suggest that a decrease in pitch between phrases shows some sort of topic subordination and hence groups phrases together, whereas an increase in pitch signals a new unit of some sort, such as a new topic or a new turn. To investigate these predictions and test them against my spontaneous and read speech data, I constructed hierarchical pitch trees. These trees were based on high pitch heads which dominate lower pitch phrases. These trees were specifically designed to capture relationships between phrases in which relationships of increasing pitch between phrases work to divide discourse into different segments and relationships of decreasing pitch between phrases signal coherence between the phrases. That is, if pitch range increases at new topic boundaries and new turns, these boundaries should be captured by a division into separate trees. On the other hand, if an increase in pitch range is used for other purposes besides marking boundaries between discourse segments, we would not expect these trees to capture those relationships clearly. For example, if certain kinds of relationships between phrases are made by increasing pitch from one

phrase to the next instead of by decreasing pitch, the grouping predicted by the trees based on decreasing relationships would be a mismatch with what should be grouped together. The pitch trees impose a segmentation upon the discourses, which I called the pitch tree segmentation.

I considered these trees to be a kind of phonetic structure which captures in a gradient way which phrases are grouped together by decreasing pitch relationships. No *a priori* categories of pitch ranges (e.g. low, mid, high) were assumed. However, these could be assigned later if such a categorization seemed appropriate (see Bruce and Touati, 1992, for work which uses such a categorization). The phonetic structure can be interpreted later, much as a fundamental frequency contour is a phonetic representation which can be interpreted phonologically in terms of accents and phrases. If rising pitch relationships between phrases are uncovered as well, then clearly a richer structure which captures increasing as well as decreasing relationships is called for.

Hierarchical pitch trees were constructed from the peak pitch values of each intermediate phrase in a text (peak measurement criteria as described in Section 6.1). The peak F0 of each intermediate phrase was taken as an acoustic measure of the pitch range for each phrase. This algorithm built hierarchical pitch trees based on the principle that a high pitch dominated all subsequent lower pitches until the next local increase. That is, phrases with subsequently decreasing pitch ranges were grouped together, and phrases where pitch range increased were divided into separate groups. The first higher pitch value in a sequence started a new group. Three levels of groupings were constructed. The first level of trees, Level 1, was based on the measured peak of each phrase. The next two levels, Level 2 and Level 3, were based upon the highest values of each tree in the immediately lower level. The value of the highest daughter became the value used for building the next level of the tree. So, the values for Level 2 were the highest values from the level-1 trees, and the values for Level 3 were the highest values from the level-2 trees. The nested levels of trees captured the large increases in pitch appropriate for changes in topic and the like.

Let us illustrate the step-by-step construction of a hierarchical pitch tree using this algorithm with the example given in Fig. 12, an excerpt which was taken from the read version of FP's conversation. At the left are the intermediate phrases of the text; the underlined words are the accented words which have the peak F0s. The column labeled Peak F0 shows the F0 measured in Hz for the underlined words. The trees at each level begin with a frequency value which is higher than

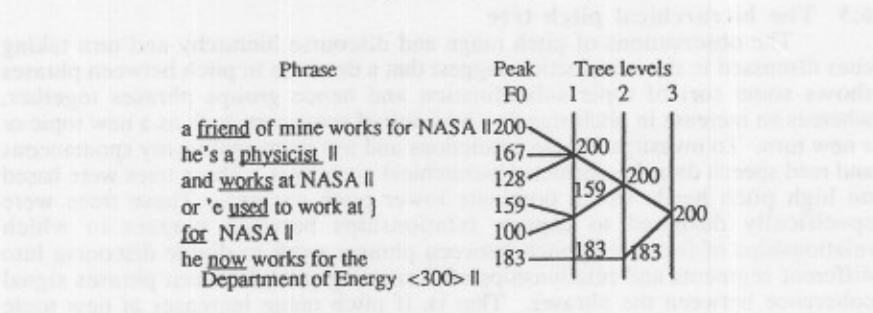


Fig. 12. Building a hierarchical pitch tree. The underlined words are the accented words on which the peak F0 for each phrase was realized. The peak F0 value for the phrase is given in Hz in the column Peak F0. Trees were constructed from these values according to the algorithm described in the text. From Speaker FP, Read: 'Friend'.

the following value. The first level-1 tree begins with the first node, which has a value of 200 Hz. This node dominates the next two nodes, which are realized on progressively lower frequencies. Note that the trees need not be binary branching. A new tree on any level begins at a local increase in frequency values. For example, the second level-1 tree begins with a node of 159 (the peak on *used*), which is greater than the previous node of 128. The 159 is greater than the next node value of 100, so the two nodes are together in a tree. Similarly, the third level-1 tree begins with the node value of 183 (for *now*) because 183 is greater than the previous node of 100. This tree happens to be a tree with only one branch. Levels 2 and 3 proceed similarly, with the value of the highest daughter becoming the value for the next level. So, the first level-2 tree dominates the first two level-1 trees, which have values of 200 and 159 respectively. The second level-2 tree is again the single branch dominating 183. At level 3 there is a single tree which dominates the inherited values of 200 and 183.

### 7. Comparing discourse and pitch tree segmentations

The trees presented in Figs. 13-18 are schematic but represent the full structure of the pitch segmentation trees for selected parts of Speaker FP's conversation and Speaker DW's conversation presented in Figs. 1-6. Triangles represent selected full trees at the three different levels, neglecting the internal structure of the trees. The pitch values that head the trees are circled in the figures. Heavy lines show the divisions into "paragraphs" that the coder marked. As in Figs. 1-6 my utterances are shown in shaded boxes. The relationships marked by arrows labeled 'C' and 'T' show rising pitch relationships for corrections marked by the coder and what I am calling introductory phrases (see below). We will look at the spontaneous and read versions of two different sections of Speaker FP's conversation ('College' shown in Figs. 13 and 14, and 'Friend' shown in Figs. 15 and 16) and one section of Speaker DW's conversation ('Fernblaster' shown in Figs. 17 and 18).

I will be looking at these data specifically to test my hypotheses that the pitch range cues to topic structure in the read speech versions conform fairly well to what has been suggested, i.e. that pitch range is expanded at the beginning of new topics and decreased for related subtopics, but that pitch range conveys the hierarchical discourse structure of the spontaneous speech versions less well. I expect to find that real-time production phenomena such as floor negotiations, corrections, and false starts disrupt clear topic organization. These may complicate the role pitch range plays as a cue to topic structure, because they themselves may have manifestations in the pitch ranges used. That is, there should be differences between the two versions in how well the pitch range reflects the discourse structure because the read speech versions were reorganizations of the spontaneous speech versions (since they came from the same texts), ones which were free of the complexity of floor negotiations (since the turns were predefined) and false starts (since those were removed in the preparation of the texts).

Recall from Section 3 that the discourse segmentations labeled by the coder differed substantially between the spontaneous and read versions in 'College' for Speaker FP. We will see that the pitch tree segmentations also differed substantially between the two versions, and in fact matched the discourse segmentations quite well. However, for Speaker DW, neither the discourse segmentations nor the pitch tree segmentation for the spontaneous and read versions differed dramatically. The discourse segmentations suggest a substantial reorganization of the topic structure from the spontaneous speech to read speech version in FP's conversation, but a considerably lesser reorganization in DW's conversation, and these differences between the two speakers seem to be reflected in comparable relationships between the pitch trees of the paired versions of text.

## 7.1 Speaker FP

### 7.1.1 'College'

Figures 13 and 14 show the spontaneous and read versions of the 'College' part of FP's conversation. It is clear that the major divisions into paragraphs in

these two versions are different, and this is a reflection of the fact that different things were emphasized in the two versions. In the spontaneous version many of the subpoints flowed from one to the other (witnessed by the lack of paragraph breaks), while in the read version some of the transitions between points were abrupt enough for the coder to assign them to separate paragraphs. An examination of the pitch trees associated with the read version shows that each of paragraphs 2, 3, and 4 have their own pitch trees. They were all headed by pitch ranges with values of approximately 200 Hz. There were also relatively long pauses, from 434 ms to 802 ms, at these boundaries between the paragraphs. Recall from Fig. 1 that the mean pause duration in the read version was 322 ms, so these are well over the mean. It seems reasonable to assume then that the combination of a regular pause and a large increase of pitch are cues to a strong discourse boundary. This is exactly what Hirschberg and Grosz (1992) found in their AP news reading.

We can also see in the read version the tendency for a fairly hierarchical structure indicated by decreasing pitch range for supporting details of the paragraphs. For example, in paragraph 2, lower level trees headed by local increases seem to correspond nicely to the supporting details. He knew he wasn't going to be a cartographer and registered for Spanish (200 Hz). Because he was taking Spanish, the advisor suggested introductory linguistics (152 Hz) which was a course like our 201 (127 Hz). Then there is another fact, that is, Dr. Thomas Field taught the course (147 Hz). Paragraph 3 has a similarly nice structure with decrease pitch ranges for the subpoints, with the exception that the third level-2 tree peak (156 Hz) has a larger value than the second level-2 tree (133 Hz), but it is still less than the level-3 head (198 Hz).

The most striking difference between the topic structure of the two versions is this part concerning Thomas Field. In the spontaneous version, the speaker mentioned Thomas Field as the instructor of the course without making a strong point of emphasizing who he is, whereas the read version specially highlighted Thomas Field. In the spontaneous version, the pitch tree for the section mentioning him as the instructor of the course had a peak of 127 Hz, and the first mention of Thomas Field had a peak value of 125 Hz. He mentioned Thomas Field a second time along with his title, also in a low pitch range (peak 111 Hz), and then that Mary knows him because he went to Cornell. He elaborated essentially as a parenthetical that Thomas Field graduated from Cornell and not only went there. The only accent on the phrase *graduated from Cornell* was on *graduated*, and this was realized with a peak of 108 Hz. The whole phrase was uttered with low intensity and at a relatively fast pace. All of the vowels which were not devoiced or deleted had z-score durations between 0 and -1. All of these cues together make the parenthetical type meaning of *graduated from Cornell* clear to the listener, and the coder assigned no paragraph breaks separating the discussion of Thomas Field from the previous material. However, the read version specially highlighted Thomas Field, and the discourse segmentation and the pitch tree segmentation both reflect this. The rise in the pitch range (from 139 Hz to 198 Hz) and the pause between the first and second mention of Thomas Field in the read version signaled a clear separation from the previous mention of him as the instructor of the course. A separate paragraph was devoted to him, and one of the points made about him was that he graduated from Cornell. In this version of the phrase *graduated from Cornell* the standardized vowel durations were all over 0, and the final vowel of *Cornell* had a z-score of 2, with the word accent on *Cornell* and phrase final



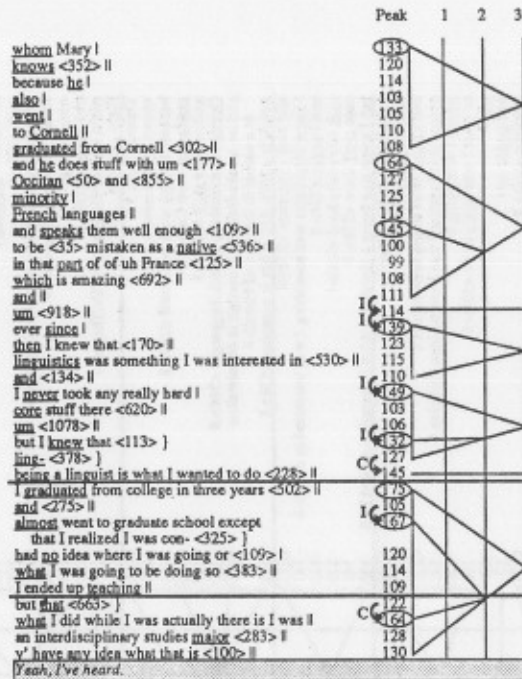
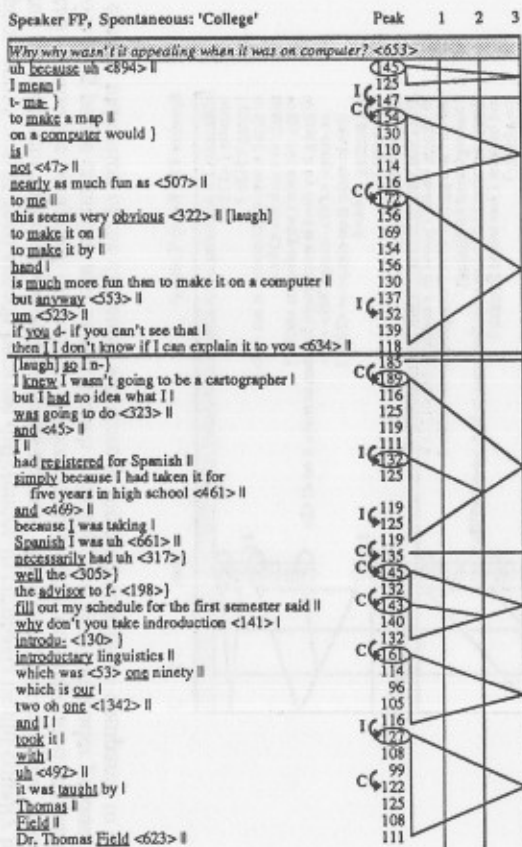


Fig. 13. Hierarchical pitch trees.  
Speaker FP, Spontaneous: 'College'

lengthening contributing to the long length. A paraphrase of the meaning of this part is something like the following. The man who Mary knows because she went to Cornell with him, graduated from Cornell. The listener can tell the difference in meaning between the two versions quite easily due to the pitch range relationships, pause structure, and tempo.



These differences in emphasis are not solely due to a difference in spontaneous versus read speech because differences in emphasis might equally be true for different instances of spontaneous speech. However, the fact that in the read version the paragraph boundaries marked by the coder corresponded regularly with long pause durations and extreme pitch rises suggests that the reader produced a clearer indication of the discourse structure in the read version as compared to the

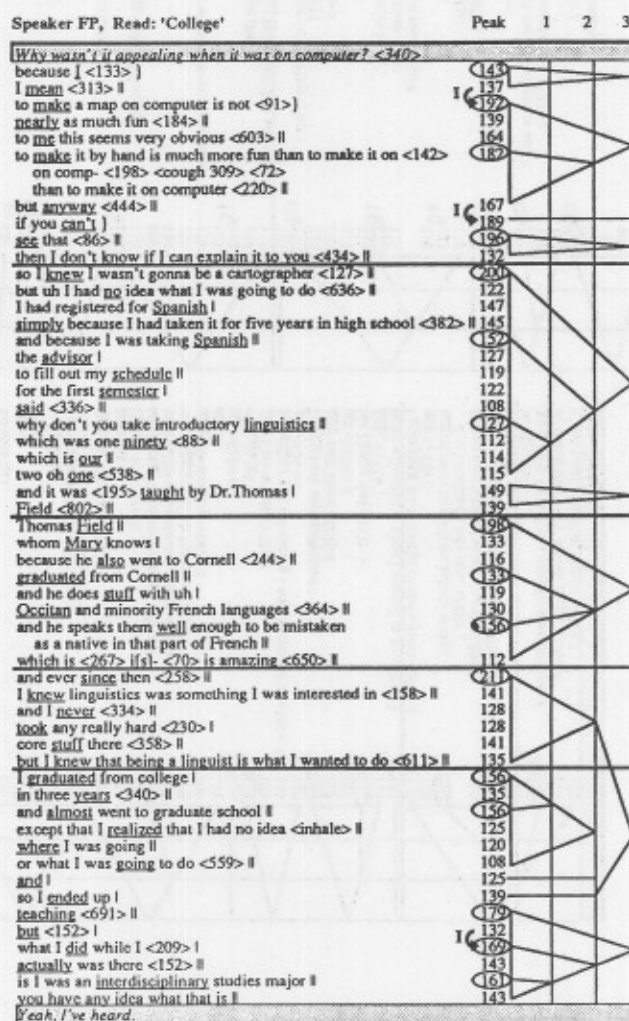


Fig. 14. Hierarchical pitch trees. Speaker FP, Read: 'College'

spontaneous version. I would claim that this is an instance of exactly the kind of reorganization and simplification of discourse structure that I expected to find between the spontaneous and read versions.

### 7.1.2 'Friend'

Figures 15 and 16 show the spontaneous and read versions of the 'Friend' part of FP's conversation. The first paragraph of the spontaneous version ends with the sentence which was given as example (8) above, where the high rising intonation was an indication that the speaker expected verbal feedback from me. However, I didn't give it to him, and he continued speaking after an extremely long pause of 1250 ms. He made a few false starts before he started speaking fluently again. The trees in this section after the long pause were headed by increasing pitches: 149, 159, 164. It seemed that FP started out at one pitch range and increased the pitch with each subsequent attempt at topic, a kind of topic reset. The phrases grouped together by the first pitch tree after the pause (headed by 149 Hz) turned out to be a false start which ended with the phrase *the stuff he knows is kind of amazing 'cause*. This phrase had a peak of 128 Hz and a following 1137 ms pause. The next phrase *he does a lot of* was a new attempt after that false start. It had a peak of 159 Hz, which was higher than that of the immediately previous phrase and was also higher than the peak of the whole group which included that phrase. The tree headed by 164, the highest of all the peaks in this section, seems to be his main point, that his friend knows a lot of things. Notice that the tree at level 3 headed by 164 spans a pause gap of 1196 ms after a complete unit, suggesting another possible turn transition point. After the pause, FP raised pitch locally (from 116 to 141 Hz) again as a mark of starting a new topic or a new turn, but not as high or higher than 164 Hz, the peak of the section to which it was topically related. French and Local (1986) note that pitch is raised for competitive turn taking. These data suggest that pitch is also raised (or reset) after a point when a turn could have taken place even when the other speaker did not compete for a turn. Such a turn transition point is an appropriate place to either provide more information as a subtopic or elaboration of the previous topic, and thus make a smaller rise in pitch, or to suggest a new topic, and thus make a larger rise in pitch.

The topic structure of the last part of the read speech, corresponding to the second paragraph in the spontaneous version, seemed to be something like this. The main topic of discussion was the stuff the friend knows. He knows more than physics; specifically he knows geography and climate. *Geography* and *climate* were relatively more prominent than *physics* (realized with a peak of 169 Hz on geography as opposed to 147 Hz on physics), but they were both examples of what he knows. Instead of simply having hierarchical subtopics, this section had levels of parallelism expressed in the pitch ranges. The two versions shared the topic organization that geography and climate are examples of things that he knows. However, in the read version they were given in explicit comparison with physics, whereas in the spontaneous version they seem to have been details added partly because I did not take the floor.

### 7.1.3 Introductory phrases

The correspondence between the auditory discourse segmentation and the pitch tree segmentation are nearly identical in the read version of the section 'College'. Furthermore, the predicted relationship of decreasing pitch range with subtopic structure seemed to hold fairly well in the read versions. In the spontaneous section 'Friend' an interesting connection with possible turn transition points and pitch trees were shown. However, while the pitch trees grouped phrases together into paragraphs quite well, they did not always group phrases of sentences together correctly. One specific type of situation where the pitch tree

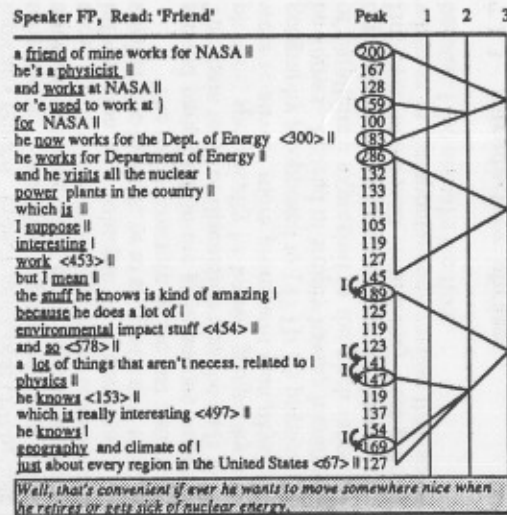
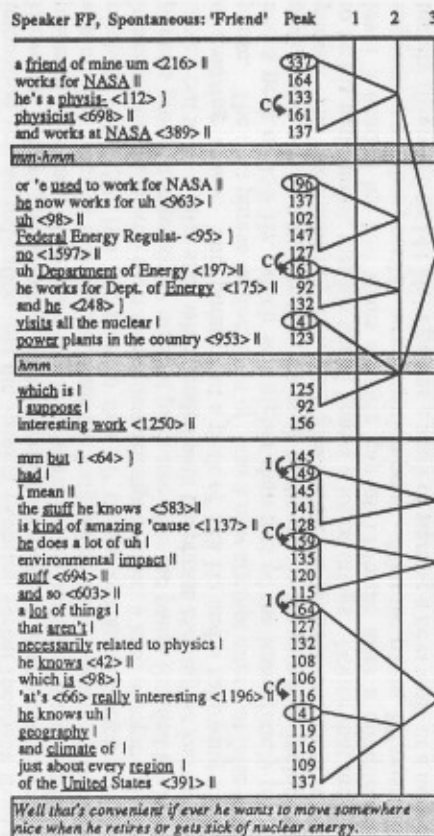


Fig. 15. Hierarchical pitch trees.  
Speaker FP, Spontaneous: 'Friend'

Fig. 16. Hierarchical pitch trees.  
Speaker FP, Read: 'Friend'

algorithm made wrong predictions about which phrases to group together were sentences which did not start out with a phrase realized in the highest pitch range for the sentence. Introductory phrases such as *and*, *and so*, *um*, *but I mean*, and the like, were generally realized in lower pitch ranges than the more content-containing part of the sentence. These were more common in the spontaneous speech than the read speech, but they occurred in both versions. The tree building algorithm as it was defined did not group such introductory phrases together with the following phrase as it should have, but rather grouped them with a previous phrase which had a higher pitch range. Examples of this sort are marked with arrows labeled 'I' (for 'introductory') in Figures 13 to 16. This sort of relationship between phrases seems to be related to the intonation of cue phrases. Cue phrases introducing a phrase are often realized with a L\* accent, and they are therefore realized at a lower frequency than the phrase that they are in relationship to (Hirschberg and Litman, 1987). These introductory phrases were realized with H\* accents and not L\*, but the increasing pitch relation between an introductory phrase and a following larger discourse unit was similar.

#### 7.1.4 Corrections

The pitch tree segmentations did not correspond as well to the auditory discourse segmentation in the spontaneous versions as they did in the read versions. One thing that is apparent from the spontaneous speech versions is that they were full of false starts, corrections to mispronunciations of words, and irregularly distributed pauses of various lengths. In stark contrast to the read versions, the spontaneous versions were riddled with such reflections of the unprepared nature of the text. The speaker did not know ahead of time what he was going to say, and had to create it on-line as he spoke. Sometimes the speaker made mistakes and had to correct what he said to what he intended. Each of the arrows marked with 'C' (for 'correction') were places where FP aborted a false start and started anew or repeated a word in a new phrase as a correction. Compare the spontaneous versions (Figs. 13 and 15), which had many corrections between phrases with the read versions (Figs. 14 and 16), which did not have such corrections. Recall that the discourse codings discussed in Section 3 listed corrections for the read version. These corrections were of a different nature, however, with the corrections being corrections for the most part being within the same phrase and simple repetitions of a word which was stumbled over in reading. In the spontaneous speech, a correction was almost always uttered with a higher pitch than the word or phrase corrected. That is, there was a local increase in pitch between the phrases. There were a few examples of corrections between phrases being uttered on a lower pitch, such as the *went to Cornell, graduated from Cornell* example, but these were parenthetical additions of information and do not feel like true corrections.

There were several examples of such increases in pitch range for false starts. The spontaneous example described in Section 7.1.2. 'Friend' was such an example. The increase from 122 to 164 Hz in the false start sequence *but that* <663> *what I did while I was actually there is I was* in the last paragraph of 'College' is another example of such an increasing relationship in false starts. The false start in the first paragraph of 'College' ending with the phrases *is not nearly as much fun as* <507> which was aborted and then corrected by a new approach beginning with the phrase *to me* also had an increasing pitch relationship. The last phrase of the false start had a peak of 116 Hz and the new start beginning had a peak of 172 Hz. A false start reformulation *ling-* <378> *being a linguist* had an increase from 127 to 145 Hz. The peak pitch for each phrase in the string of false starts *because I was taking Spanish I was uh* <661> *necessarily had uh* <317> *well the* increased from 119 to 135 to 145.

Corrections at the level of the word also exhibit this kind of increasing pitch range relationship. The second mention, the correction, was realized on a higher pitch than the first, incorrect, mention. The correction can be due to incorrect or incomplete pronunciation the first time, such as *introdu-* <130> *introductory* with an increase from 132 to 162 Hz, *make a map* with an increase from 147 to 154 Hz, and *physic-* <112> *physicist* <678> with an increase from 133 to 161 Hz (from 'Friend'). Factual corrections also have this sort of pitch relationship, such as the example *Federal Energy Regulat-* <95> *no* <1597> *uh Department of Energy* <197> from 'Friend'. There is an increase from 147 to 161 Hz from *Federal Energy to Department of Energy*, with *no* at 127 in between. This increase of pitch range seems to be a quite general tendency and a way to mark a new beginning of a correction. These local increases for corrections however disturb the trend for hierarchical topic organization to be marked by decreasing pitch range relationships within a topic group and an increase at the beginning of a new topic group. We might view this kind of pitch increase for a correction as one cue that the listener might take advantage of in recovering the final form of what was intended, as Clark and Schaefer (1989) say listeners can.

## 7.2 Speaker DW

Figures 17 and 18 show the spontaneous and read versions of part of the 'Fernblaster' part of DW's conversation. Both the discourse segmentations and the pitch tree segmentations were quite similar for the spontaneous and read speech versions. One mismatch between the trees and the discourse segmentation was the division between the second and third paragraphs. The division between the second and the third paragraph did not align neatly with the pitch trees in either version. However, for the spontaneous version there was a 436 ms pause at the end of the second paragraph, and the next few phrases could be taken as introductory phrases to a new point. For the read version there was a short pause of 103 ms at that boundary and a local pitch increase from 110 to 143 Hz.

The rest of the pitch trees were quite similar in the two versions. In the first paragraph the tree was headed by 159 Hz in the spontaneous and 156 Hz in the read. In the second paragraph there were a few trees, which were headed by quite high peaks on the phrases *it's a weird sounding name*, *Fernblaster*, and *and your eighth grade English teacher* in both versions. Essentially he was role playing and quoting himself and his students, and he used the same sort of changes in pitch range to signal that in both versions. His background comments were uttered with smaller ranges, with peaks of less than 115 Hz, such as the phrases *and you hear these little titters in the back of the room*. The fourth paragraph began with almost exactly the same peak pitch in both versions (neglecting the introductory phrase in the spontaneous version), 159 Hz for the spontaneous and 161 Hz for the read version. So, not only did the pitch tree segmentations essentially match in the two versions of this part of the conversation, the values of the peaks were also nearly identical, signaling parallel emphasis in the two versions. In both the spontaneous and read versions Speaker DW was re-enacting the scene from his class by quoting himself and his students, partly by use of high pitch ranges in the quoted phrases. Using pitch range for quoting in this way disrupted hierarchical topic structure but revealed very similar use of pitch ranges in the two versions.

Just as for Speaker FP, Speaker DW had examples of introductory phrases that were not grouped with the following phrase by the pitch tree as they should have been, but instead with the previous phrases in both the spontaneous and the read versions. Again these are marked in the figures with arrows labeled by 'I' for introductory. The read version had no examples of corrections, but the spontaneous version did, and they are marked with arrows labeled by 'C' for correction in the figure. He repeated the word *people* in the two subsequent

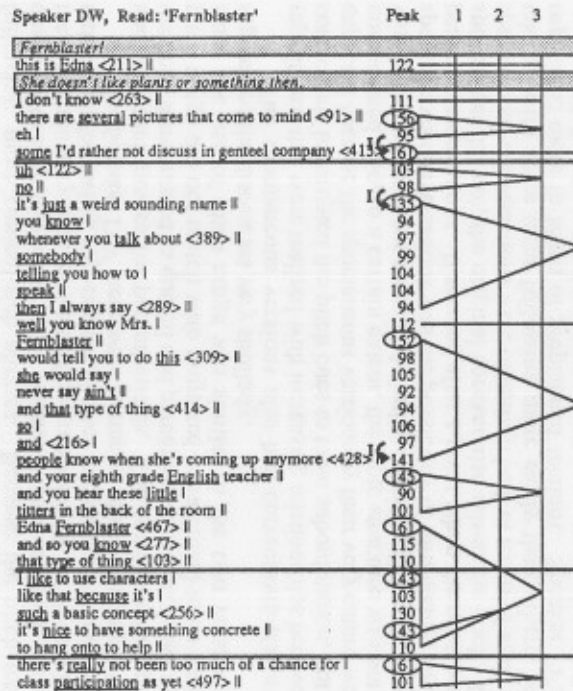
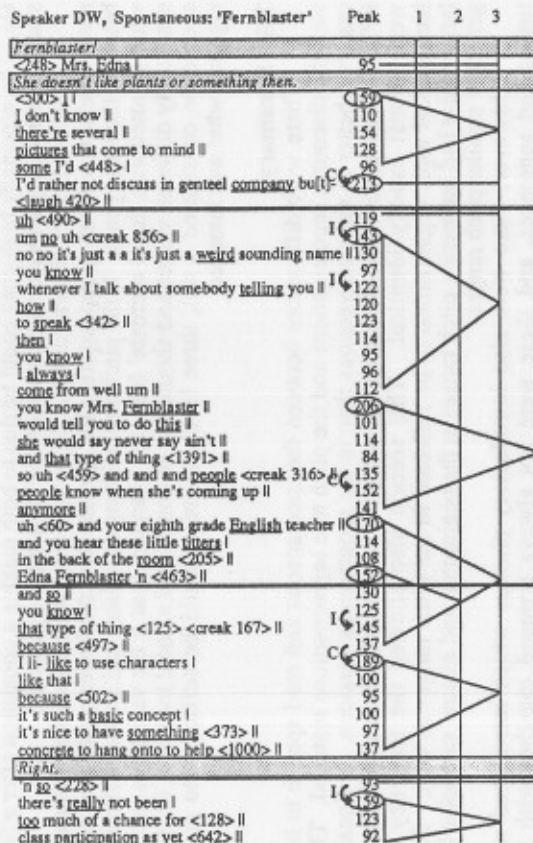


Fig. 17. Hierarchical pitch trees. Speaker DW, Spontaneous: 'Fernblaster'

Fig. 18. Hierarchical pitch trees. Speaker DW, Read: 'Fernblaster'

phrases *so uh* <459> *and and and* *people* <316> *people know when she's coming up* with an increase of pitch from 135 to 152 Hz from the first to the second mention. He corrected a false start beginning with *because* by starting up again after a 502 ms pause with the phrase *I li- like to use characters* with a peak of 189 Hz. The especially high peak could be due to both making a correction after a false start and correcting a mispronunciation of *like* by a second mention.

The pitch trees shown for this section had trees with four levels instead of three levels in order to group together the phrases with the lower peaks that occur between the large peaks. If the pitch tree building algorithm had another criteria of what counted as a 'local increase' (say, for example, that 5 Hz variations are not essentially different values and should not be considered a 'local increase', i.e. they should be considered a tie), three levels of trees would be sufficient to group together what was intended.

### 7.3 Summary

There were differences between the spontaneous and read speech in how well the discourse segmentations and the pitch tree segmentations matched. They matched better in the read versions than in the spontaneous versions. In Speaker FP's 'College' section, the discourse segmentation and the pitch tree segmentation were almost exactly identical. This section also showed the tendency for hierarchical topic organization to be reflected in the pitch ranges as well. New large topics had increasing pitch ranges at the beginning, and related subtopics had generally smaller pitch ranges.

However, there were also introductory phrases which introduced new topics and sentences, and these were not always grouped together with the appropriate phrases. These phrases were realized with lower pitch range than the following, more content-rich phrases. Because the pitch tree algorithm was designed to treat increasing pitch relationships as division points between units, these introductory phrases were often grouped with previous, higher pitch phrases rather than following, higher pitch phrases as they should have been.

For Speaker DW there were examples of using high pitch ranges for quotes and low pitch ranges for parentheticals. The corresponding phrases in the two versions were treated as quoted and parenthetical material, but again these uses of pitch range did not match with projected hierarchical topic structure. However, since the use of pitch range was similar in the two versions, the pitch trees segmented the discourses very similarly.

The spontaneous versions also had corrections (false starts and word repetitions) that were realized with increasing relationships between phrases. These corrections interrupted the pitch cues to topic subordination, but corrections were expected because the spontaneous versions by their very nature were unplanned and unrehearsed. So, to a certain extent, the topic structure was not as clear to begin with in the spontaneous speech. A further complication to the topic structure in the spontaneous conversations was the possibility of turn taking, since these were two person conversations. It seemed after a point when the other speaker could have spoken but did not, the original speaker also raised the pitch. I propose that the spontaneous conversations were organized both in terms of topic structure and turn taking, with some turns following more easily than others, and the read were organized more in terms of preplanned sections. Speaker FP's read version of section 'College' was a clear example of a reorganization of the contents of the spontaneous version. It was a reading made with knowledge of what was coming up next and how long each turn was to be and without hesitations, false starts, and other corrections.

The discourse segmentations were quite different for Speaker FP between the spontaneous and read versions of the same conversations, even though the



words were nearly identical in the two versions. This means that things were grouped together differently and given different emphasis in the two versions. Pitch range relationships did reflect the differences in discourse structure. The paragraph boundaries in FP's read version corresponded to regular pauses and quite large pitch range expansions, and in FP's spontaneous version again it was at the points of largest pitch range expansions that paragraph divisions were marked. However, the discourse segmentations were nearly identical in the two versions for Speaker DW. We could interpret this as saying that Speaker FP changed the topic relationships more between the spontaneous and read versions than Speaker DW did. Speaker DW seemed to have more or less preserved the organization of the original spontaneous conversation, judging from the discourse segmentations and similar use of pitch range.

## 8. Perception test

The materials used in this study differed from the spontaneous versus read speech used in such studies as Remez et al., 1985; Remez et al., 1986; Blaauw, 1991; Blaauw, 1992. Since the read speech was a connected discourse based on spontaneous speech and was deliberately read with the aim of trying to make it sound spontaneous, I wondered how well the readers had succeeded in their task. That is, was the read speech perceived as spontaneous or read? In addition, the two speakers differed dramatically in the extent to which the pitch range patterns reflecting topic organization corresponded between the spontaneous and read versions of the conversations. For Speaker FP the two versions were very different, while for Speaker DW they were very similar. I wondered if these differences between speakers could be partially explained by characteristics of the read speech versions. Specifically, did DW remember or recreate the structure of the spontaneous speech in his read version more so than FP did in his (as the use of pitch range would lead us to believe), and if so, was Speaker DW's read speech more spontaneous sounding than Speaker FP's? If this were true, then we would expect to find excerpts from DW's read version perceived as spontaneous more often than excerpts from FP's read version would be perceived as spontaneous. Finally, I wondered if longer excerpts were more often correctly identified as spontaneous or read than shorter excerpts. Longer excerpts may contain more cues to the spontaneous or read nature of the text than shorter excerpts because the larger amount of material is more likely to contain hesitation phenomena in the spontaneous, etc., and may reveal to the listener differences in patterns of transitions between phrases and topics in the two modes of speech. To address these questions, I designed a perception test to test how well listeners could correctly identify excerpts of these spoken conversations and the reenacted read speech as spontaneous or read. Listeners were presented with utterances from each speaker and different lengths of utterances.

### 8.1 Method

*Subjects.* Twenty eight undergraduate linguistics students volunteered to participate in the experiment. All were native speakers of American English and none reported any hearing impairment.

*Stimuli.* The spontaneous and read conversations of both speakers were segmented into utterances one sentence long, three sentences long, and five sentences long. The three sentence and five sentence long utterances overlapped by one sentence at the beginning and one at the end with other members of the series. Thus each of the single sentences occurred at least once and at most twice in the three sentence utterance set and the five sentence utterance set.

*Design and procedure.* A stimulus tape consisting of two parts was prepared, and items were presented in blocks of 10, with a three second interstimuli

interval. Listeners could take a break between the two parts. Part I included 110 one sentence long utterances in random order (2 speakers x (29 + 26)), and Part II included 78 three and five sentence long utterances in random order (2 speakers x ((13+13) 3 sentences + (7+6) 5 sentences). There were 28 listeners and 188 items,

for a total of 5264 responses.

Listeners sat in a soundproof room listening to the stimulus tape over headphones and for each item circled either 'spoken' or 'read' on an answer sheet. 'Spoken' meant they thought the excerpt they heard could have come from a naturally occurring conversation between two friends, and 'read' meant that they thought that the excerpt they heard came from a reenactment of a conversation, read from a transcript of a naturally occurring conversation. They were told that the readers were trying to make the reading sound as much like a spontaneous conversation as possible, so that it might be difficult to tell whether it was spontaneous or read. They were told they were there to judge how well the readers had done in reading naturally. The task took approximately 40 minutes.

## 8.2 Results

Chi-squared tests showed that there was a significant effect of speaker on perception of the utterances as spontaneous or read. More of DW's utterances were perceived as spontaneous than were FP's ( $\chi^2(1) = 159.14, p < .01$ ). Listeners perceived 68% of DW's utterances as spoken and 51% of FP's as spoken. In actuality, half were spontaneous and half were read for each speaker. Fig. 19 shows these judgments in the columns labeled 'perceived as spontaneous' and 'perceived as read'. The columns labeled 'ss' are the spontaneous utterances which were perceived as spontaneous, and 'rs' are the read utterances which were misperceived as spontaneous. The columns labeled 'rr' are the read utterances which were perceived as read, and 'sr' are the spontaneous utterances which were misperceived as read. The three different shaded columns in each of the categories

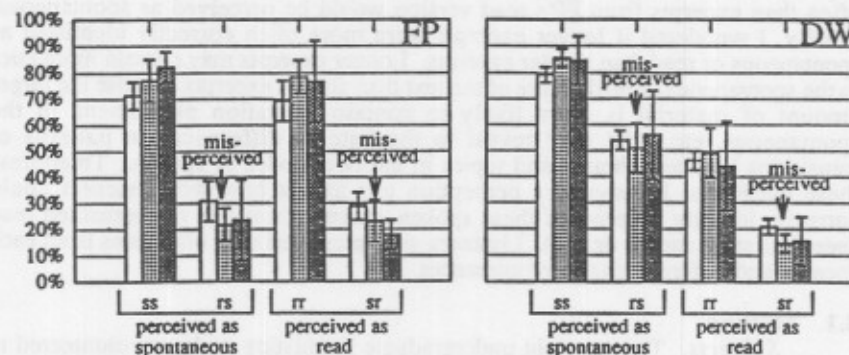


Fig. 19. Perception results in judging material spontaneous or read for the two speakers. (left) Speaker FP, (right) Speaker DW. Many more of DW's read utterances were misperceived as spontaneous utterances than were FP's.

ss: spontaneous perceived as read, rs: read perceived as spontaneous, rr: read perceived as read, sr: spontaneous perceived as read. Three different lengths of material are shown: white fill is 1 sentence long, medium fill is 3 sentences long, darkest fill is 5 sentences long. 95% confidence intervals are marked.

show the three different lengths of utterances. The white fill represents 1 sentence long utterances, the medium fill 3 sentence long utterances, and the darkest fill 5 sentence long utterances. As we can see from the columns labeled 'rs', many more of DW's read utterances were misperceived as spontaneous than FP's. We also see that fewer of DW's spontaneous utterances are misperceived as read as compared to FP. The percentage perceived correctly as spontaneous or read was significantly different for the speakers, 73% for FP and 64% for DW ( $\chi^2(1) = 51.9, p < .01$ ).

For Speaker FP, length of the utterance (i.e. 1 sentence, 3 sentences, or 5 sentences) had a significant overall effect on the number of correct judgments made (i.e. spoken when spoken, read when read), ( $\chi^2(2) = 27.2, p < 0.01$ ). Further analysis revealed that the difference was significant only for the shortest (1) vs. longer (3, 5) utterances (1 vs. 3:  $\chi^2 = 17.3, p < 0.01$ ; 3 vs. 5:  $\chi^2 = 0.24, p > 0.1$ ). Longer utterances were judged correctly more often. This could be because more turn taking clues are provided in the longer utterances. However, for Speaker DW, length of utterance had no significant overall effect on the number of correct judgments ( $\chi^2(2) = 3.612, p > 0.05$ ). That means that there was no significant difference for the shortest vs. longer utterances for this speaker. Utterances were no more likely to be perceived correctly when they were longer.

### 8.3 Discussion

The perception test revealed that there were significant differences between the two speakers as to how the speech materials were perceived. Speaker FP's utterances were judged correctly on average about 73% of the time, with more of the longer utterances (three and five sentences) judged correctly than the shortest (one sentence) utterances. However, only 64% of Speaker DW's utterances were judged correctly, and there was no significant effect of length of utterance. More of DW's read utterances were misperceived as spontaneous than were correctly perceived as read. This was not true for Speaker FP. This lends support to the interpretation that DW read more naturally than FP and succeeded in producing a read text that sounded quite spontaneous. The fact that the shorter utterances were more often misjudged than the longer utterances for Speaker FP could mean that longer excerpts from the conversations or readings gave the listeners more clues to the true mode of speech. However, for Speaker DW the listeners did not categorize longer excerpts correctly more often, perhaps meaning that DW succeeded in reenacting spontaneous relationships between phrases in the read speech.

The results of these listening tasks are clearly at odds with the claims that listeners know immediately whether they are listening to natural spontaneous speech or read speech. Perhaps it would be more realistic to say that people do not really know whether an utterance is spontaneous or read, but that they make judgments early. For example, in a gating experiment with Dutch (Blaauw, 1992), listeners were able to classify utterances as spontaneous or read about 82% of the time when given the full sentence, but as well as 63% given the first two syllables and 75% given the first 6 syllables. Since in my experiment listeners were only correct on average 73% or 64% depending on the speaker given full sentences and even several sentences together, we must say that the differences between spontaneous and read speech are not as clear-cut as they might at first seem. It seems more likely that there is a continuum between clearly spontaneous and clearly read speech, with differences in style being quite important. Hesitations, false starts, long pauses and the like are prototypical of spontaneous speech, but spontaneous speech does not have to be disfluent. Read speech is often syntactically distinct since it is based on written texts. Since the read speech in this task was based on spontaneous speech, the syntax was more typical of spontaneous

speech than a read text. The additional instruction to the readers to read the transcript to make it sound like a spontaneous conversation further blurred the edges between the two kinds of speech.

Perhaps the comparison between the two speech styles in this study would benefit from being considered in terms of the scales unprepared versus prepared or unrehearsed versus rehearsed. On those scales, Speaker FP's spontaneous narrative was less prepared than DW's spontaneous narrative, because DW has spoken about his teaching methods and the use of the character Mrs. Fernblaster before. We have talked about teaching experiences together before, and have had conversations discussing pedagogical methods. So, with more rehearsal still, DW's read version is not likely to change its organizational structure as much as FP's read version of a story which he had not told before. There are also individual differences between people and their acting ability, and hence how well they can reenact a conversation and make it seem natural.

## 9. Discussion

Two different spontaneous conversations were recorded and reenacted as read speech by the original speakers. A listening test involving categorizing excerpts from these conversations as spontaneous or read showed that accurate identification was not entirely straightforward. Many of the read utterances were perceived as spontaneous, and some of the spontaneous utterances were perceived as read. Many more of Speaker DW's read utterances were perceived as spontaneous than Speaker FP's. Apparently skilled readers reading material based on spontaneous conversation can succeed to a certain extent in producing utterances that sound convincingly spontaneous.

The patterns of results for the two speakers were not identical, which is not particularly surprising, given that the listening test determined that the two speakers succeeded to different degrees in producing read speech that sounded like spontaneous speech. Several acoustic measures were made to see if they distinguished the two versions. Pause duration measures revealed that both speakers had similar pause duration distributions, with a higher mean and larger standard deviation of pause duration in the spontaneous than in the read speech. These results match previous findings. A measure of fundamental frequency, the mean F0 peak per phrase, distinguished Speaker FP's spontaneous from read speech, but it did not distinguish Speaker DW's spontaneous and read speech. Measures of average F0 and F0 range have found different relationships depending on language and the specific materials used; this speaker difference is another such result.

A symbolic phonological intonational analysis found some consistent patterns in the differences between spontaneous and read speech. The phrases in the read version were longer on average than the phrases in the spontaneous version. The transcription also showed that there was no use of the H-H% high rising contour as grounding or checking to see that the listener understood. It seems that there was interaction with the listener in the spontaneous version which was missing in the read version. The read speech lacked the hallmarks of interactivity in the spontaneous speech except the ones that are inherent to the text (change of speaker, explicit questions). We could say then that the read version was more like coordinated monologues rather than a true dialogue. The read version was like the spontaneous minus true interaction between the speakers.

The discourse organizations were clearly different between the spontaneous and read versions for Speaker FP, but they were relatively similar for Speaker DW. The pitch tree algorithm (based on measures of the peak pitches of all intermediate phrases) provided a method for comparing the organizational structure of matched spontaneous and read speech discourses. It provided a way of testing the

predictions about how pitch range is used to signal topic structure. The segmentation that the pitch tree algorithm imposed upon the discourses corresponded quite closely to the discourse segmentation that the independent coder assigned to the discourses. The best match between the pitch tree segmentation and the discourse segmentation was in the read version of Speaker FP's section 'College'.

Although the spontaneous and read versions were nearly identical in terms of syntax, different items were marked as salient, and topics were grouped together differently. The read versions were grouped into sections with relatively clear hierarchical topic structures. The spontaneous versions showed some evidence of hierarchical topic structure, but they also had disruptions to these topic organizations due to false starts, corrections, and the influence of possible turns. I hypothesize that the planned production units differ between spontaneous and read speech. I propose that spontaneous conversations are organized both in terms of topic structure and turn taking, with some turns following more easily than others, and read conversations are organized more in terms of preplanned sections. In the read versions, the readers know exactly what is coming up and do not have to negotiate for turns with the conversational partners. This gives them more control over deciding what relationship to give to the various topics. One meaning of pitch increase seems to be a reflection of the start of a new unit, whether it is a new topic or a new turn.

The pitch tree segmentations, together with the discourse annotations, showed that pitch increased in these discourses at the beginning of new topics and at the beginnings of new turns, or potential turns. This matches previous findings. Each such pitch range increase started a new hierarchical tree of descending pitch. The pitch tree algorithm relied on these pitch increases to segment the text, and so could only capture relationships among phrases such as topic subordination and sentence internal declination. However, there were also relationships among phrases based on increasing pitch, for example, false starts, corrections, and introductory phrases. These relationships could only be represented indirectly in the descending pitch trees built by the algorithm. The pitch trees helped to explore the multifunctional use of pitch range changes without first having to posit categories of pitch range and abstract away from the phonetic signal.

The pitch tree algorithm for representing pitch trees could benefit from some fine tuning. As I have defined it now, any local increase in pitch gives rise to a new pitch tree at the appropriate level. Very small differences in frequency, such as 1 to 5 Hz should probably not count as differences in level. Such small differences can be due to measurement errors or inherent fundamental frequencies of different vowels and probably are not even reliably distinguished by listeners. Further work would need to be done to determine how big a difference should be represented as a difference, and if it depends on the absolute location in the frequency range. However, it has been interesting to see how much could be learned by using this extremely simple coding of the conventional wisdom that pitch increases for new topics and that subtopics have pitch ranges less than their main topics. The method showed that this was true to a certain extent in even quite complicated texts, spontaneous as well as read, but that this was not the whole story. It revealed a need to be able to represent connective increasing pitch relationships as well decreasing pitch relationships for such things as the possibility of introducing a new topic and subsequent corrections.

#### References

- Beckman, M.E. & Pierrehumbert, J.B. (1986) Intonational structure in Japanese and English, *Phonology Yearbook*, 3, 255-309.

- Blaauw, E. (1991) Phonetic characteristics of spontaneous and read-aloud speech. In *Proceedings, ESCA Workshop on Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication*, pp. 12/1-12/5. Barcelona.
- Blaauw, E. (1992) On the perceptual difference between read and spontaneous speech: Two experiments, *OTS Yearbook 1992* (M. Everaert, B. Schouten & W. Zonneveld, editors), pp. 1-16. Utrecht, The Netherlands: LED.
- Bolinger, D. (1978) Intonation across languages. In *Universals of Human Language, 2* (Phonology) (J. H. Greenberg, editor), pp. 471-524. Stanford University Press.
- Brazil, D., Coulthard, M. and Johns, C. (1980) *Discourse intonation and language teaching*. Longman.
- Brown, G., Currie, K. L. & Kenworthy, J. (1980) *Questions of Intonation*. London: Croom Helm.
- Brown, G. (1983) Prosodic structure and the given/new distinction. In *Prosody: Models and Measurements* (D. R. Ladd and A. Cutler, editors), pp. 67-68. Berlin: Springer-Verlag.
- Bruce, G. & Touati, P. (1992) On the analysis of prosody in spontaneous speech with exemplification from Swedish and French, *Speech Communication, 11*, 453-458.
- Butterworth, B. (1975) Hesitation and semantic planning in speech, *Journal of Psycholinguistic Research, 4*, 75-87.
- Campbell, W. N., and Isard, S. D. (1991) Segment durations in a syllable frame, *Journal of Phonetics, 19*(1), 37-47.
- Campbell, W. N. (1992) Prosodic encoding of English speech. In *Proceedings, Second International Conference on Spoken Language Processing, 1*, pp. 663-666. Banff, Canada.
- Clark, H. H., and Schaefer, E. F. (1989) Contributing to discourse, *Cognitive Science, 13*, 259-294.
- Cooper, W. E. & Paccia-Cooper, J. (1980) *Syntax and Speech*. Harvard University Press.
- French, P. & Local, J. (1986) Prosodic features and the management of interruptions. In *Intonation in discourse* (C. Johns-Lewis, editor), pp. 157-180. San Diego, CA: College-Hill Press, Inc.
- Grosz, B. J. & Sidner, C. L. (1986) Attention, intentions, and the structure of discourse, *Computational Linguistics, 12*(3), 175-204.
- Grosz, B., and Hirschberg, J. (1992) Some intonational characteristics of discourse structure. In *Proceedings, Second International Conference on Spoken Language Processing, 1*, pp. 429-432. Banff, Canada.
- Gårding, E. (1967) Prosodiska drag i spontant och uppläst tal. In *Svenskt talspråk* (G. Holm, editor), pp. 40-85. Uppsala, Sweden: Almqvist & Wiksells Boktryckeri AB.
- Hirschberg, J. & Litman, D. (1987) Now let's talk about now: Identifying cue phrases intonationally. In *Proceedings, 25th Annual Meeting of the Association for Computational Linguistics*, pp. 163-171. Stanford, CA.
- Hirschberg, J. & Grosz, B. (1992) Intonational features of local and global discourse structure. In *Proceedings, Fifth DARPA Workshop on Speech and Natural Language*, pp. 441-446. Harriman, NY: Morgan Kaufmann.
- Hirschberg, J. & Pierrehumbert, J. (1986) The intonational structuring of discourse. In *Proceedings, 24th Annual Meeting of the Association for Computational Linguistics*, pp. 136-144. New York.
- Howell, P. & Kadi-Hanifi, K. (1991) Comparison of prosodic properties between read and spontaneous speech material, *Speech Communication, 10*, 163-169.

- International Phonetics Association (1989) Report on the 1989 Kiel Convention, *Journal of the International Phonetics Association*, 19(2), 67-80.
- Ladd, D. R. (1993) Constraints on the gradient variability of pitch range. In *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III* (Patricia Keating, editor), pp. 43-63. Cambridge University Press.
- Lehiste, I. (1975) The phonetic structure of paragraphs. In *Structure and Process in Speech Perception* (A. Cohen and S. G. Nooteboom, editors), pp. 195-203. Springer-Verlag.
- Lehiste, I. (1979) Perception of sentence and paragraph boundaries. In *Frontiers of Speech Research* (B. Lindblom and S. Ohman, editors), pp. 191-201. London: Academic Press.
- Lehiste, I. (1980) Phonetic characteristics of discourse. Presented at the Meeting of the Committee on Speech Research, Acoustical Society of Japan.
- Levin, H., Schaffer, C. A. & Snow, C. (1982) The prosodic and paralinguistic features of reading and telling stories, *Language and Speech*, 25, 43-54.
- Lieberman, M. & Pierrehumbert, J. (1984) Intonational invariance under changes in pitch range and length. In *Language Sound Structure: Studies in phonology* (M. Aranoff and R. T. Oehrle, editors), pp. 157-233. MIT Press.
- Passenout, R. J. & Litman, D. J. (1993) Feasibility of automated discourse segmentation. In *Proceedings, 31st Annual Meeting of the Association for Computational Linguistics*, pp. 148-155. Ohio State University.
- Pierrehumbert, J. B. (1980) *The phonology and phonetics of English intonation*. Doctoral dissertation, Massachusetts Institute of Technology, September 1980. Distributed by the Indiana University Linguistics Club.
- Pierrehumbert, J. & Hirschberg, J. (1990) The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication* (P. R. Cohen, J. Morgan & M. E. Pollack, editors), pp. 271-311. MIT Press.
- Remez, R. E., Rubin, P. E. & Ball, S. (1985) Sentence intonation in spontaneous utterances and fluently spoken text. Presented at the 109th Meeting of the Acoustical Society of America, Austin, TX, April 1985.
- Remez, R. E., Rubin, P. E. & Nygaard, L. C. (1986) On spontaneous speech and fluently spoken text: Production differences and perceptual distinctions. Presented at the 111th Meeting of the Acoustical Society of America, Cleveland, OH, May 1986.
- Sacks, H. & Schegloff, E. A. (1979) Two preferences in the organization of reference to persons in conversation and their interaction. In *Everyday Language: Studies in ethnomethodology* (G. Psathas, editor), pp. 15-21. New York: Irvington Publishers.
- Schegloff, E. A. (1982) Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In *32nd Georgetown University Roundtable on Languages and Linguistics 1981, Analyzing discourse: Text and talk*, (D. Tannen, editor), pp. 71-93. Washington, DC: Georgetown University Press.
- Shockey, L. R. (1974) Phonetic and phonological properties of read speech, *Ohio State University Working Papers in Linguistics*, 17, iv-143.
- Silverman, K. E. A. (1987) *The structure and processing of fundamental frequency contours*. Doctoral dissertation, University of Cambridge.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992) TOBI: A standard for labeling English prosody. In *Proceedings, Second International Conference on Spoken Language Processing*, 2, pp. 867-870. Banff, Canada.

## When is a Syllable not a Syllable?\*

Mary E. Beckman

mbeckman@ling.ohio-state.edu

**Abstract:** This paper reviews evidence for unifying two seemingly disparate types of syllable reduction phenomena: the elision of reduced vowels in English and German, and the devoicing of high vowels in Japanese, Korean, and Montreal French. Both types of "casual speech rule" can be understood as extreme endpoints of a phonetic continuum of gestural overlap. The vowel is seemingly deleted or devoiced when the gestures of neighboring consonants encroach so completely into the space for the affected vowel that the relevant vowel gesture(s) leave no salient acoustic trace. However, in some cases in some of these languages, the reduction has been phonologically reanalyzed, so that the word loses a syllable. The paper explores the circumstances under which such reanalysis can occur.

### 1. Introduction

Recent work on the gestural organization of speech (e.g., Browman & Goldstein 1990a, 1990b) supports alternative phonetic accounts of such segmental reduction phenomena as the devoicing of high vowels in Japanese and Korean and the deletion of schwa and simplification of consonant clusters in English and German. Whereas earlier phonological descriptions assumed these to be categorical changes of phonological form, akin to the alternations seen in the inflectional morphology of the same languages, we now can describe them as byproducts of subtle shifts in the articulatory specifications of gestural magnitude and timing — shifts which can cause dramatic changes in the acoustic realization of a particular segmental string because of nonlinearities in the mapping between the two phonetic representations. For example, in fast-speech productions of German *mit dem Wagen* [mit<sup>h</sup> dɐm v'a:gən] 'by car', the temporal distance between the oral gestures for the consonants [d] and [m] in *dem* and between those for [g] and [n] in *Wagen* might be reduced to the point of aerodynamically hiding the release of the alveolar or velar stop, thus effectively deleting the unstressed vowels, as in [mit<sup>h</sup> dɐŋ v'a:gŋ] (Kohler 1990). The devoicing of the first [w] in Japanese /supootu/ [supo:tsu] 'sports' might be

---

\*This paper was originally presented at the Dokkyo International Forum on Speech Recognition and Phonology, Dokkyo University, Soka City, Saitama Prefecture, Japan, 18-19 December 1993, and will appear in the annual report of the Dokkyo University International Center. As the many citations of "personal communication" suggest, I am indebted to Louise Levac for sharing with me her intuitions as a native speaker of Montreal French after her talk at the ESCA Workshop on Prosody in Lund in September 1993, and to Stefanie Jannedy for her extreme generosity with her as yet unpublished data and her astute observations of her native language. The analysis of Korean vowel devoicing owes to Sun-Ah Jun much more than is suggested by the simple reference to Jun & Beckman (1993). I thank Beth Hume, Mariko Kondo, Kikuo Maekawa, and Arnold Zwicky for giving me copies of several papers to which I otherwise would not have had access.



described as a similar hiding of the vowel's oral and glottal gestures by those of the preceding fricative (Jun & Beckman 1993).

From the point of view of segmental production and perception, these reduction phenomena are strikingly similar across languages. The acoustic consequence of gestural overlap between the first consonant and vowel in /supootu/ is virtually indistinguishable from the consequence of extreme gestural overlap in English *supports*; both result in something that English speakers readily perceive as an initial [sp] cluster. However, the prosodic consequences of these segmental reductions can be strikingly different. In Japanese, the syllable count is necessarily preserved even under such complete reduction that no acoustic trace of vowel formants remain, and /supootu/ will always have three syllables, no matter whether the first vowel is devoiced or not. In English, by contrast, the comparable reduction of an unstressed vowel can in effect delete the syllable, so that the disyllable *supports* can be confused with the monosyllable *sports*. These different consequences highlight the cognitive robustness of such prosodic units as the syllable, and the very different roles that such units play in speech rhythms across languages.

This paper will survey the phonological circumstances under which syllable count can change as a result of extreme vowel reduction. Since a thorough review of all languages in which such reduction phenomena have been observed would be impossible, the bulk of the discussion will be limited to only five languages, with a view to representing a wide range of rhythmic types — mora-timed Japanese, syllable-timed French and Korean, and stress-timed English and German. The paper will begin by describing the phonetics of vowel reduction in more detail, reviewing some of the evidence in the literature supporting the newer phonetic accounts of the phenomenon which differentiates it from phonological processes that yield superficially similar alternations.

## 2. Deletion of /ə/ in English and German

In English and German, the prosodically weakest syllables — i.e., unstressed syllables with an underlying /ə/ vowel — often undergo even further reduction, to the point of effectively losing any salient phonetic trace of the vocalic nucleus. Table 1 illustrates the segmental conditions conducive to this effect, without differentiating among the varying speech rates or styles where some of these are likely to occur (e.g., [ʔəfli] for *awfully* is readily available even in more careful, formal speech contexts where [k<sup>h</sup>æfɪpm] for *captain* would be produced only as a deliberate rusticism). In both languages, the /ə/ is very likely to be “deleted” in this way when it is flanked either by an obstruent and a sonorant or by an obstruent and a sibilant fricative. In medial syllables, this segmental context allows an apparent heterosyllabic consonant cluster to result. In initial or final syllables, the sonorant or fricative consonant apparently replaces the vowel as the syllabic nucleus. In German, it is not uncommon also to “delete” the vowel between two sonorants, a context where English is more likely to retain some acoustic remnant of the underlying vowel (as can be seen by comparing the speech rates or styles in which the pronunciations given in the table for *Kannen* and *linen* are likely). The last block of the table illustrates that the phenomenon is not necessarily limited to underlying /ə/, but can apply to other lax vowels in comparably weak prosodic position.

Superficially, these alternations between forms with and without a vowel resemble morphological alternations such as the variants of the regular possessive form in English (e.g., *Pat's* /pæts/, *Bob's* /bɒbz/, *horse's* /hɔːsɪz/, but *Horace's* /hɔːrəsɪz/ ~ *Horace'* /hɔːrəs/, *Achilles's* /əkɪlɪzɪz/ ~ *Achilles'* /əkɪlɪz/). And that is how they are treated in many earlier descriptions of them. For example, Zwicky

(1972) describes the pronunciations listed in Table 1 for English words such as *mystery* and *happening* in terms of "Slur", a phonological rule that deletes the reduced vowel (i.e., [ə] → Ø) between any consonant and an unstressed syllable

beginning with [l], [r] or [n].<sup>1</sup> Strauss (1982) similarly describes German as

having optional phonological rules of reduction to schwa and schwa deletion to account both for the weak forms of function words, such as [dəm] ~ [d̩] for *dem* [de:m] cited above, and for casual-speech productions of content words such as [ʔa:d̩] for *Adel* and [laɪt̩] for *leiten*. Hall (1992) differentiates these two types by describing the latter set of pronunciations as the result of the non-application of an optional "Schwa-Default Rule" in his schwa-epenthesis account of forms such as [ʔa:d̩] and [laɪt̩], but nonetheless makes them both categorical phonological processes.

Table 1. Schwa-deletion (and unstressed /ɪ/- or /ε/-deletion) in English and German.

English		German	
<i>beret</i>	/bə're/ [bje]	<i>beraten</i>	'advise' /bə's'a:tən/ [bɛʔa:t̩]
<i>collapse</i>	/kəl'æps/ [k <sup>h</sup> læps]	<i>geleiten</i>	'accompany' /gəl'aɪtən/ [glaɪt̩]
<i>Toledo</i>	/təl'ido/ [t̩'ɪro]		
<i>suspect (v)</i>	/səsp'ekt/ [s̩sp <sup>h</sup> ekt]	<i>subtil</i>	'subtle' /zʊbt'ɪl/ [ʃp̩t̩'l]
<i>support</i>	/səp'ɔ:t/ [sp <sup>h</sup> ɔ:t]		
<i>cotton</i>	/k'atən/ [k <sup>h</sup> at̩n]	<i>leiten</i>	'lead' /l'aɪtən/ [laɪt̩]
<i>sudden</i>	/s'ʌdən/ [sʌd̩n]	<i>leiden</i>	'suffer' /l'aɪdən/ [laɪd̩]
		<i>Glauben</i>	'faith' /gl'aʊbən/ [glaʊb̩]
<i>captain</i>	/k'æptən/ [k <sup>h</sup> æp̩t̩n]	<i>fettem</i>	'greasy (dat)' /f'etəm/ [fep̩]
<i>linen</i>	/l'ɪnən/ [lɪn̩]	<i>Kannen</i>	'pitchers' /k'anən/ [k <sup>h</sup> an̩]
<i>pommel</i>	/p'ʌməl/ [p <sup>h</sup> ʌm̩]	<i>Himmel</i>	'sky' /hɪməl/ [hɪm̩]
<i>bottle</i>	/b'atəl/ [bət̩]	<i>Adel</i>	'nobility' /a:dəl/ [ʔa:d̩]
<i>mystery</i>	/m'ɪstəri/ [m'ɪst̩]	<i>gelegene</i>	'located (f)' /gel'egənə/ [gel'eg̩nə]
<i>coordinate</i>	/kə'ɔ:dɪnət/ [k <sup>h</sup> ɔ:d̩nɪt]	<i>geladene</i>	'invited (f)' /gel'adənə/ [gəl'ad̩nə]
<i>happening</i>	/h'æpənɪŋ/ [h'æp̩nɪŋ]	<i>Ebenen</i>	'plateaus' /e:bənən/ [ʔe:b̩nən]
<i>awfully</i>	/ɔ'fəli/ [ʔɔ'f̩li]	<i>adelige</i>	'noble (f)' /a:dəlɪgə/ [ʔa:d̩lɪgə]
<i>horoscope</i>	/h'ɔ:skɒp/ [h'ɔ:sk̩ɒp]	<i>Horoskop</i>	'horoscope' /hɔ:skɒp/ [hɔ:sk̩ɒp]
<i>Morris</i>	/m'ɔ:ɪs/ [mɔ:ɪ̩]	<i>wahres</i>	'true (n)' /v'a:ɪs/ [va:ɪ̩]
<i>synopsis</i>	/sɪn'ɒpsɪs/ [ʃnɒps̩]	<i>Symbol</i>	'symbol' /zɪmb'ɔ:l/ [zɪmb̩'ɔ:l]
<i>symbolic</i>	/sɪmb'ɒlɪk/ [sɪmb̩'ɒlɪk]	<i>Schimpanse</i>	'chimpanzee' /ʃɪmp'anzə/ [ʃɪmp̩'anzə]
<i>Chicago</i>	/ʃɪk'ɑ:go/ [ʃk'ɑ:go]	<i>Schikane</i>	'annoyance' /ʃɪk'a:nə/ [ʃk <sup>h</sup> 'a:nə]
<i>vicinity</i>	/vɪs'ɪnɪti/ [v̩s'ɪnɪti]		

Kohler (1990), by contrast, argues against all such traditionally generative phonological accounts. If it is described by symbolic rules of the sort posited by Strauss (1982) and Hall (1992), the apparent deletion of the unstressed vowel in weak forms of monosyllabic function words and in various other weak syllables in

<sup>1</sup>"Slur" is a separate rule from the "Pre-stress Contraction" that Zwicky posits for forms such as [v̩s'ɪnɪti] for *vicinity* and [k<sup>h</sup>læps] for *collapse*. For the latter set of forms, he considers that they may be "merely automatic consequences of faster speech", although he points out that such an account "presumes a coherent and detailed theory of linguistic phonetics, which [in 1972] cannot be said to be available yet" [Zwicky 1972, p. 278].

German is difficult to relate to the consonant assimilations and substitutions that tend to coöccur with the schwa deletion in connected or casual speech. For example, the phrase *mit dem Wagen* 'by car' might be realized as any of the segment strings in Table 2, showing the output at each step in the application of possible rules in a text-to-speech system. These rules can be grouped into sets for progressively more casual stylistic modules. However, while it might be useful for text-to-speech synthesis systems to discretize the stylistic progression in this way, such rule modules cannot provide an explanatory account of what speakers are actually doing in producing the apparent progression of forms transcribed in the table. On the other hand, if each closely related set of transcribed feature changes is instead understood as the acoustic byproduct of some articulatory restructuring in the interest of "motor economy", the seemingly disparate changes that tend to coöccur within a "rule module" can be explained and predicted in terms of the phonetic process involved. For example, when stated in terms of phonological rules, the progression from [gən] to [ŋ] involves two discrete and formally unrelated changes: schwa deletion and nasal place assimilation. When understood as a phonetic reorganization, however, it might be stated as a single articulatory change: letting the gesture for the dorso-velar constriction extend well into the vowel so that the magnitude of its releasing phase is drastically reduced could simultaneously "delete" the [ə] and aerodynamically "hide" any more anterior apical constriction gesture. Kohler describes many other similar examples in support of the idea that formulating a precise phonetic model of articulatory timing and gestural magnitude could unify schwa deletion with the stylistically related consonant assimilations and weakenings.

**Table 2.** Realizations of *mit dem Wagen* 'by car' after the application of each applicable casual-speech rule in a German text-to-speech system. (Kohler, 1990.)

[mit <sup>h</sup> de:m v'a:gən]	input form
[mit <sup>h</sup> ɸe:m v'a:gən]	devoicing of voiced stops after voiceless consonants
[mit <sup>h</sup> ɸəm v'a:gən]	vowel reduction in weak form of function word
[mit <sup>h</sup> ɸm v'a:gən]	schwa deletion before nasals after stressed syllable
[mit ɸm v'a:gən]	deaspiration of voiceless stops before stops and nasals
[mɪp ɸm v'a:gən]	regressive place assimilation of apical nasals and stops
[mɪp ɸm v'a:ŋ]	progressive place assimilation of apical nasals and stops
[mɪ ɸm v'a:ŋ]	degemination of devoiced or voiceless consonants
[mɪ b m v'a:ŋ]	voicing of plosives in unstressed function words
[mɪ m m v'a:ŋ]	regressive nasal assimilation of voiced stops
[mɪ m v'a:ŋ]	degemination of other consonants

Browman and Goldstein (1990a, 1990b) propose such a phonetic model. The "gestural score" represents an utterance as a set of discrete dynamically-specified control regimes for accomplishing such basic tasks as forming a constriction somewhere in the vocal tract. These control regimes (or "gestures") are temporally coördinated within and across parallel channels, with each channel (or "tier") allocated to the control of a different articulatory subsystem. Thus, for example, the initial /bər/ sequence in the English word *beret* is represented in the score as a set of labial-closing and rounding gestures on the lip tier, an apico-postalveolar approximation gesture on the tongue-tip tier, and so on. The approximation gesture on the tongue-tip tier can be timed to overlap somewhat with the labial-closing and releasing gesture on the lip tier. The audible presence or apparent deletion of the [ə] can be modelled simply by changing the degree of overlap between the constriction gestures on the two tiers. Indeed, because the gestural-score representation has been implemented as the front end to an articulatory synthesis system, Browman

and Goldstein (1990b) could generate a continuum of degrees of gestural overlap in a series of stimuli which they presented to listeners for identification. The identification function shows a category shift, with subjects perceiving *beret* for the stimuli with least overlap, but *bray* for the stimuli with greatest overlap. This result

is in keeping with Price's (1980) earlier study showing that the difference between *parade* and *prayed* or *polite* and *plight* can be synthesized by varying such things as the resonant consonant's duration and amplitude in ways that mimic the acoustic results of varying gestural overlap.

The generalization from such results to the representation of actual fast-speech productions is suggested already in Browman and Goldstein's (1990a) seminal paper on the gestural score. If the apparent deletion of the vowel nucleus and neutralization with the corresponding consonant cluster is demonstrably nothing more than the extreme endpoint of an attested continuum of degrees of reduction and confusability, then the continuously variable values of overlap in the gestural score is a better representation than a categorical phonological rule of schwa deletion for fast-speech pronunciations of words such as *parade*, *support*, *sudden*, and *happening*. Further support for this idea comes from data in Manuel et al. (1992), who document subtle acoustic cues suggestive of an underlying "hidden" glottal-adduction gesture in *support* even in tokens that could be transcribed as [spɔʊt] and misperceived as *sport* (see also Fokes & Bond, 1993).

Jannedy (1993) gives evidence for an analogous gestural overlap account for apparent schwa deletion in German. She had northern German subjects produce a paragraph in which were embedded, in segmentally and prosodically similar contexts, words from minimal pairs such as *braten* (/bʁ'atən/ 'fry') versus *beraten* (/bɛʁ'atən/ 'advise') and *Kannen* (/k'anən/ 'pitchers') versus *kann* (/kan/ 'be able to'). She had the subjects read the paragraph ten times, starting first at a comfortable "neutral" rate and then producing four repetitions in progressively faster versions, then returning to neutral rate to produce a series of five more repetitions at progressively slower tempi. She then excised the target words from context and presented them to native speakers for identification in a forced-choice judgment. Her results for *Kannen* versus *kann* are particularly illuminating. Plotting the duration of the target /nən/ in *Kannen* or /n/ in *kann* against the duration of the remaining /ka/ in the production of *kann* (used as a metric of the overall speech rate), she found that the two regression functions converged at the fastest rates. Moreover there was a comparable convergence in the identification functions: at faster and faster rates, more and more of the listeners misjudged *Kannen* to be *kann*. For *beraten* versus *braten*, there was less convergence in the regression lines for the production data, but as much in the perception data, albeit in the other direction: listeners often misjudged *braten* to be *beraten* at the slower rates. Moreover, in no case was there evidence of a bimodal distribution in the duration or identification measures — thus, no evidence of a categorical shift from presence to absence (or from absence to presence) of a reduced vowel phone. Jannedy concludes that German has neither a schwa-deletion rule, as proposed by Strauss (1982), nor a schwa-insertion rule, as proposed by Hall (1992). That is, the apparent alternation between forms with and forms without schwa is an artificial imposition of two symbolic categories onto a continuum of degrees of encroachment by the neighboring consonants' gestures onto those of the vowel.

### 3. "Devocalization" of high vowels in Japanese, Korean, and Montreal French

Jun and Beckman (1993) propose a comparable gestural-overlap representation for another common reduction phenomenon involving the devoicing or deletion of high vowels. Such devocalization has been studied most extensively for standard (Tokyo) Japanese, but it also occurs in the Montreal variety of French and in at least

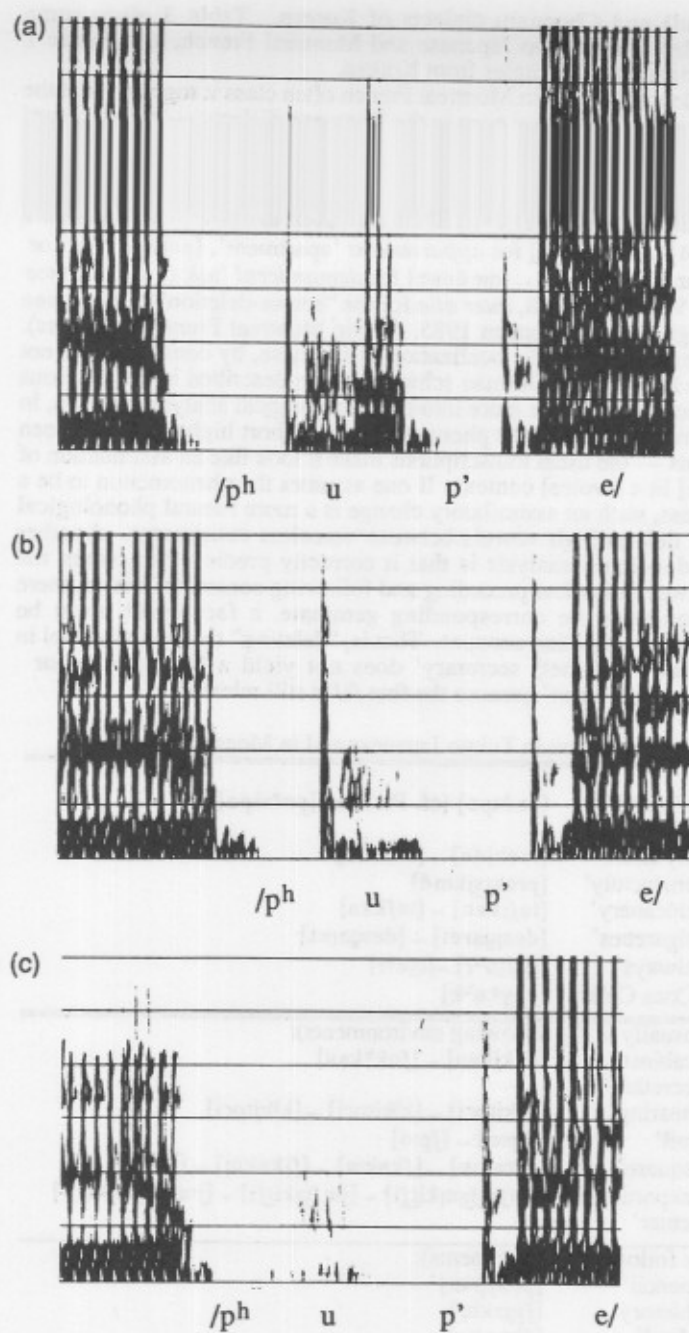
the standard (Seoul) and Chonnam dialects of Korean. Table 3 gives some transcribed examples from Tokyo Japanese and Montreal French, and Figure 1 shows some spectrograms of examples from Korean.

Descriptions of devocalization in Montreal French often class it together with the alternations involving [ə] that occur even in the European dialects — e.g., standard Parisian [apʁɛtəmɑ̃] ~ [apʁɛtmɑ̃] for *appartement* 'apartment', [pʁɛti] ~ [pti] for *petit* 'small (m)', or [dɑmɑ̃dɑ̃vɛ] ~ [dmɑ̃dvɛ] for *demandera* 'ask (1 s, fut)' (see Delattre 1951, and Verluysen 1988, *inter alia* for the "schwa-deletion" of European French, and Cedergren and Simoneau 1985, for the Montreal French processes). Most traditional descriptions of devocalization in Japanese, by contrast, have not equated it with the English and German schwa-deletion described in the previous section, perhaps because there is a more intuitive phonological analysis. That is, in the most typical environment for the phenomenon — a short high vowel between voiceless consonants — the usual transcriptions make it look like an assimilation of [+voice] to [-voice] in a [-voice] context. If one assumes the phenomenon to be a phonological process, such an assimilatory change is a more natural phonological rule than one that deletes high vowels between voiceless consonants. Another advantage of the devoicing analysis is that it correctly predicts that, when the process occurs between identical preceding and following consonant onsets, there is no neutralization with the corresponding geminate, a fact which must be stipulated separately in a deletion account. That is, "deleting" the medial vowel in /sjokikan/ [ʃok(i)kan] '(cabinet) secretary' does not yield a homophone for /sjokkan/ [ʃok:an] 'tactile organ' because the first /k/ is still released.

Table 3. Devocalization in Tokyo Japanese and in Montreal French.

Montreal French:		
<i>principaux</i>	'principal (pl)'	[pʁɛʃpo] (cf. Parisian [pʁɛʃipo])
<i>mes idées</i>	'my ideas'	[mezide] ~ [mezde]
<i>pratiquement</i>	'practically'	[pʁaktʃjkmɑ̃]
<i>la chicane</i>	'chicanery'	[laʃikan] ~ [laʃkan]
<i>des cigarettes</i>	'cigarettes'	[desjɛgʁɛt] ~ [desgʁɛt]
<i>toujours</i>	'always'	[tʃuʒuʁ] ~ [tʃuʁ]
<i>du Coke</i>	'Coca Cola'	[dzykoʔk]
Tokyo Japanese (usually in the following environments):		
/sjoki'kan/	'cabinet secretary'	[ʃokjkan] ~ [ʃokʰkan]
/kikitori/	'hearing'	[kjkitori] ~ [kʰkitori] ~ [kʰktori]
/sippo/	'tail'	[ʃjp:o] ~ [ʃp:o]
/si'kaku/	'square'	[ʃjakaw] ~ [ʃkakaw] ~ [ʃjakaw] ~ [ʃkakʰ]
/jusjutuki'ti/	'exporting center'	[jwʃwtʃukitʃi] ~ [jwʃtʃukitʃi] ~ [jwʃwtʃukitʃi] etc.
(more rarely in the following environments):		
/pe'nsiru/	'pencil'	[penʃiɾw]
/sigaku/	'history'	[ʃjgaku]
/sima'tta/	'drat!'	[ʃimat:a]
/kasutera/	'pound cake'	[kʌʃutera] (here, the first vowel is the rare case)

However, there are two aspects of devocalization in all three of these languages that suggest that the same general phonetic mechanism is at work as in German and English "schwa deletion". First, the phenomenon seems to apply "gradiently",



**Figure 1.** Tokens of Korean /pʰu p'e/ 'unripe pear' showing (a) "voiced", (b) "partially devoiced" and (c) "completely devoiced" types for the /u/ in the target syllable. The preceding context is an /e/. From Jun & Beckman (1993).

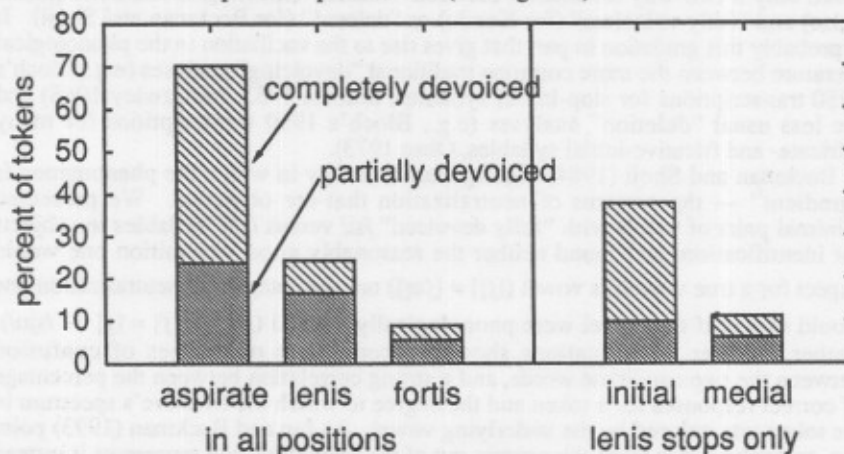
creating tokens with varying degrees of devocalization even for the same input type. For example, in their study of Montreal French high vowels, Cedergren and Simoneau (1985) could distinguish among mere "reduction" (a very short vowel — less than 30ms — but nonetheless showing some periodic energy at the second formant), "devoicing" (a somewhat shorter vowel showing only aperiodic noise excitation), and "true syncope". A similar gradation can be seen in the Korean examples in Figure 1, which illustrates the three-way classification used in Jun and Beckman (1993). We called a vowel "voiced" only when the periodic energy from the laryngeal source was large enough for a long enough interval to visibly excite the second or higher formants on the spectrogram. In addition to "completely devoiced" tokens, this criterion excluded many tokens where there were one or two weak glottal pulses visible at the bottom of the spectrograph, tokens which we called "partially devoiced". For Japanese, Maekawa (1990) used similar criteria to differentiate two possible grades of "devoicing", whereas Beckman and Shoji (1984) and Kondo (1993), while noting the gradient nature of the phenomenon, make only a two-way distinction between "voiced" (showing at least one glottal pulse) and "fully voiceless" (for Kondo) or "deleted" (for Beckman and Shoji). It is probably this gradation in part that gives rise to the vacillation in the phonological literature between the more common traditional "devoicing" analyses (e.g., Bloch's 1950 transcriptions for stop-initial syllables, Shibata 1955, McCawley 1968) and the less usual "deletion" analyses (e.g., Bloch's 1950 transcriptions for many affricate- and fricative-initial syllables, Ohso 1973).

Beckman and Shoji (1984) highlight another way in which the phenomenon is "gradient" — the patterns of neutralization that are observed. We presented minimal pairs of words with "fully devoiced" /si/ versus /sju/ syllables to subjects for identification, and found neither the reasonably good recognition one would expect for a true voiceless vowel ([ʃi] ≠ [ʃu]) nor the categorical neutralization one would expect if the vowel were phonologically deleted (/si/ → [ʃ] = [ʃ] ← /sju/). Rather, listener identifications showed a continuum of degrees of confusion between the two candidate words, and a strong correlation between the percentage of correct responses for a token and the degree to which the fricative's spectrum in the token was colored by the underlying vowel. As Jun and Beckman (1993) point out, an analysis that takes the process out of the phonology and represents it instead in terms of variable degrees of overlap in the gestural score would capture both the spectral and perceptual gradation in the signal.

The second aspect of the phenomenon that supports the gestural-score representation is that it occurs to varying extent for different segmental sequences and in different prosodic positions. For example, Cedergren and Simoneau (1985) and earlier studies of devocalization in Montreal French found that it does not occur in rhythm-group final syllables, where the vowel is prosodically lengthened and less likely to be completely covered over by the neighboring consonants' gestures (see, e.g., Levac, Cedergren & Perreault 1993 for evidence concerning rhythm-group final lengthening in Montreal French). In Japanese, similarly, devocalization is limited to phonemically short vowels, and it most commonly affects the two high vowels, which are phonetically very short even when they are not devocalized. Devocalization does occur sometimes with mid and low vowels (see, e.g., Kondo 1993, Maekawa 1990), but here it is relatively rarer. The NHK (1985) pronunciation dictionary singles out particularly the syllable types /ka/, /ko/, /ha/, and /ho/ in word-initial unaccented position, where the consonant is most strongly aspirated. For either high or nonhigh vowels, devocalization is also most common between voiceless consonants, where the vowel's glottal gesture is subject to overlap from devoicing gestures on both sides. Indeed, for high vowels in this context, devocalization is quite common even in the most careful read speech. For

example, Takeda and Kuwabara (1987) found in their analysis of a large database of citation-form productions by a professional announcer that 90% of the /si/ syllables before voiceless consonants were devoiced (see also Kondo 1993).

Since Korean contrasts three types of voiceless stops and affricates, an even finer analysis of the effect of consonantal context is possible. If the gestural overlap account is correct, devoicing should be most likely to occur after an aspirated stop, where the glottal-opening gesture is largest and extends well past the oral release, and it should be least likely to occur after a fortis stop, where the glottal-opening gesture is abruptly terminated by a tense glottal adduction just before the oral release (Kagaya 1974; Hirose, Lee, & Ushijima 1974). Comparing different prosodic positions, we can also predict that devoicing should be more common at the beginning of an accentual phrase, where aspiration is longer in an aspirated stop (Jun 1990a) and the voicing from the preceding vowel is less likely to continue well into the closure for a lenis stop (Jun 1990b, 1993; Silva 1992). These predictions were borne out in our recent study of productions by three Seoul and three Chonnam speakers' productions (see Figure 2).



**Figure 2.** Distribution of "partially devoiced" and "completely devoiced" tokens after different types of preceding consonants (left), and for the vowels after lenis stops in different prosodic positions (right). From Jun & Beckman (1993).

If devoicing is an artifact of overlap and encroachment by the gestures of an adjacent consonant, we might also predict a sequence of a voiceless fricative or affricate consonant followed by high vowel to be a particularly likely environment for the effect. The prediction stems from the observation that articulatory movements into and out of a fricative constriction are often slower than movements into and out of a homorganic stop closure (Kuehn & Moll 1976). When an oral constriction that is narrow enough to give rise to frication is slowly released into the only somewhat less constricted oral configuration for a short high vowel, the oral air pressure that has built up behind the constriction cannot be vented as rapidly as it can when the consonant is a stop or when the release is into a more open vowel. Well into the vowel, aerodynamic conditions may remain more conducive to continued turbulence at the oral constriction than to vocal fold vibration, and if the vowel is short enough, it can be effectively devoiced or deleted. The prediction of greater frequency of devoicing after fricatives and affricates is borne out in Takeda and Kuwabara's (1987) and Kondo's (1993) data for Japanese, in Jun and



Beckman's (1993) data for Korean, and in Cedergren and Simoneau's (1985) data for Montreal French.

In sum, there is a wealth of literature documenting the continuously variable nature of devocalization in Japanese, Korean, and French, and the variable extent to which it occurs in different segmental environments. Both these aspects of the phenomenon strongly suggest that devocalization should be represented as the consequence of more or less subtle changes in the phonetic specification of gestural magnitude and intergestural timing in an otherwise invariant gestural score, an analysis that is essentially identical to the account presented in Section 2 for apparent schwa deletion in English and German. Thus, the two phenomena — schwa deletion and high-vowel devocalization — can both be described as gradient phonetic reductions, presenting comparable difficulties for analyses that posit disparate synchronic phonological rules to categorically delete the segment or to categorically change the vowel's voicing features.

#### 4. The potential for prosodic reanalysis

On the other hand, when viewed over longer intervals of time, such phonetic reductions can lead to categorical changes, sometimes resulting in alternations that can be described better by synchronic phonological rules. Following Ohala (1974, 1981), we can think of sound change as a grammatical reanalysis that originates in a listener's misinterpretations of the speaker's underlying phonological intent — in this case in a misinterpretation or reanalysis of the intended prosodic structure. For example, the Modern English regular *-s* plural shows an alternation between a syllabic /z/ and a nonsyllabic /s/ or /z/, the latter apparently due to a prosodic reanalysis of consonantal overlap and consequent vowel reduction in the unstressed final syllable of Old English forms such as *stanas* 'stones' and *munecas* 'monks' (Beckman, de Jong, Jun & Lee 1992). Similarly, the modern Tokyo Japanese /t/-final verb paradigms and compound words in Table 4 show an alternation between single and geminate consonant — that is, between two short open syllables and a long closed syllable — an alternation that, in the /t/-final verb paradigms at least, must have resulted from a prosodic reanalysis of a devocalized high vowel in an earlier stage of the dialect. (For an alternative hypothesis concerning the history of the alternation in the Sino-Japanese compounds with /k/, see Itô & Mester 1993.)

Table 4. Standard Japanese alternations with /t:/ and /k:/.

verb stems	non-past	past	Sino-Japanese compounds with /k:/
/kat/ 'win'	/ka.tu/	[kat:a] (= /kat.ta/)	/ga.ku/ 'study' + /ko:/ 'school' →
/mat/ 'wait'	/ma.tu/	[mat:a] (= /mat.ta/)	[gak:ɔ:] (= /gak.ko:/) 'school'
/ut/ 'beat'	/u.tu/	[ut:a] (= /ut.ta/)	/koku/ 'country' + /ka/ 'house' →
cf /kas/ 'lend'	/ka.su/	[kafta] (= /ka.si.ta/)	[kok:a] (= /kok.ka/) 'nation'

Given that small phonetic shifts in gestural organization can lead diachronically to phonological reanalysis and sound change, we should not be surprised to see evidence that reanalysis of forms that are habitually subject to vowel reduction has occurred in some cases even in contemporary grammars. And, indeed, there is such evidence. Zwicky (1972, p.283) notes that for him the fast-speech schwa-deletion rule "Slur" is "obligatory" in the second syllables of the words *camera*, *every*, *celery*, *general*, *mystery*, *chocolate*, and *family*. That is, productions of these forms with extreme encroachment by the surrounding consonants into the /ə/ have been prosodically reanalyzed (or perhaps were originally misinterpreted in acquisition) as disyllables /kæm.ɹə/, /ev.ɹi/, /sɛl.ɹi/, /dʒɛn.ɹəl/, and so on. Ramsaran, in the revised 4th edition of Gimson (1989, p. 1989) likewise says of

British English forms such as /n'ætʃjəl/ for *natural* and /t'ɛmpjəri/ for *temporary*: "Though labelled here as 'colloquial' these elisions may occur regularly within the speech of an individual, the fuller form not forming part of his idiolect." Louise Levac (personal communication) states that devocalization has similarly led to a

prosodic reanalysis of forms such as *principaux*, which for her and many other

Montreal French speakers is a disyllable /pʁɛ.spo/ (and is transcribed as such in the prosodic transcription system used for sociolinguistic databases at l'Université du Québec à Montréal — see Levac et al. 1993).

In still other cases, we see evidence of incipient reanalysis in speaker uncertainty or in vacillating judgments about syllable count. Thus, for example, Stefanie Jannedy (personal communication) states that for her, *Kannen* feels not quite disyllabic: "It's more like a syllable and a half." Moreover, her native speaker intuition about such forms varies from word to word. By contrast to *Kannen*, *leiden* is unquestionably two syllables, and *beraten* is unquestionably three. The discrepancy between her uncertainty about *Kannen* and certainty about *beraten* seems in keeping with the perception data noted above for productions by other northern German speakers (Jannedy 1993). More than half the tokens of *Kannen* were misperceived as *kann* by the majority of the listeners, whereas none of the tokens of *beraten* was misperceived as *braten* by more than 40% of the listeners. It seems that *Kannen* is closer than *beraten* is to a communal shift to becoming a homophone for the originally shorter word.

Zwicky (1972), too, describes a continuum from "obligatory" application of "Slur" in forms such as *camera* and *every* to complete failure to apply in forms such as *graciously*, *Arabic*, and *element*. That is, for him, [kæ.mə.ɹə] is at best a stilted reading pronunciation, but [gɹɛf.sli], [æ.ɹ.bɪk] and [ɛl.mɪnt] are simply wrong. Louise Levac also says that in *citation*, which has the contrasting disyllabic *station*, the [st] sequence resulting from devocalization of the [i] is not reanalyzed as an onset cluster, and in careful lab speech [ʃ.ta.sjō] has a somewhat longer [s] and thus is not categorically neutralized with [sta.sjō]. These examples suggest that, in German and English and Montreal French, reanalysis is available for some words, but not for all words in which reduction is possible.

In Japanese, on the other hand, prosodic reanalysis is unavailable across the board. Despite the examples of phonological alternation in forms such as /gaku/ in isolation versus /gak/ for the same morpheme in /gak:o/, even the most drastic devocalization of the medial vowel in /sjokikan/ '(cabinet) secretary' does not yield a homophone for /sjok:an/ 'tactile organ'. That is, although the preceding and following consonants' glottal gestures usually overlap and blend together so thoroughly that no hint of a glottal adduction for the /i/ is present, the overlap of the consonant's oral gestures is never so great as to effectively delete the release of the first stop, even in cases such as /sjokikan/, where the resulting form would be phonotactically possible in the modern language.

Thus the reinterpretation of underlying prosodic structure is a natural but not a necessary endpoint for every continuum of gestural overlap. How, then, can we explain the variability? What are the conditions conducive to a prosodic reanalysis of syllables with extremely reduced vowels? Conversely, what factors seem to be correlated with a resistance to such a prosodic reanalysis? These questions seem worth pursuing not just for their own sake, but also for what they can tell us about the relationship between physical, phonetic structures such as stop releases and vowel formants, and such cognitive, phonological categories as the syllable.

##### 5. Phonotactic constraints on the shape of the syllable

The factor that will be examined at most length in the rest of this paper has already been suggested. Prosodic reanalysis should be difficult, if not impossible, whenever the result would be a consonant cluster or a syllable type that is rare or

unattested in the rest of the lexicon. This seems to be the primary consideration blocking reanalysis of forms with devocalized vowels in Modern Japanese. The language has only one type of tautosyllabic cluster: onset clusters with /j/, as in /kʲaku/ 'visitor'. It has only three types of closed syllable: those that end in the first mora of a geminate consonant, as in /kok.ka/ 'nation'; those that end in the moraic nasal /n/, as in /kon.ki/ 'endurance'; and those very rare overlong syllables that end in both, as in /ron.donk.ko/ 'Londoner'. Thus, except for the sequences of two like consonants that constitute geminates and the sequences of moraic nasal and following consonant in words such as /konki/ and /ron.donk.ko/, there are no heterosyllabic phonological consonant clusters in the language. Therefore, if reanalysis of words with devocalized vowels is limited to forms that result in legitimate consonant clusters, then it cannot apply in the overwhelming majority of cases. Reanalysis would not be available for the devocalized vowels in [kʲikitori] 'hearing', [sʲɔ:po:tʷ] 'sports', and [ʲɪ:kakʷ] 'square' because /kk/, /sp/ and /sk/ are not possible onsets. It would not be available for [kafʲita] 'lent' and the second vowel in [kʲikitori] because /st/ and /kt/ are not possible consonant sequences even across a word-medial syllable boundary. The only cases where reanalysis would be possible in principle are those where the devocalized vowel happens to come between two like obstruents within a word — e.g., /sjokikan/ cited above, or /kasi-situ/ [kafʲitsʷ] 'rented room'. As was also noted above, however, even these cases are not reanalyzed in modern Tokyo Japanese, and native speakers take care to audibly release the first consonant in stop-stop sequences, apparently to preserve the syllable count. It is noteworthy that it is the syllable count that matters here, since mora count would be preserved even if the consonant were reanalyzed as the coda half of a geminate. This offers further support to Kubozono's (1989, forthcoming) claim that the syllable is a psychologically real unit above the mora in Japanese.

In Korean as well, very similar phonotactic constraints on consonant sequences and syllable structure conspire against reanalysis. Like Japanese, Korean has no tautosyllabic clusters other than consonants followed by glides in the onset, as in /kwa/ 'and' and /kʲəlhon/ 'marriage'. Therefore, the only potential reanalysis is in a  $C_1V_1C_2V_2C_3V_3$  sequence when  $V_2$  is devocalized and  $C_2$  could, in theory, be reinterpreted as a coda consonant to yield a phonotactically possible heterosyllabic cluster with  $C_3$ . Although Korean has far more closed syllables and more types of medial consonant cluster than does Japanese, the restrictions on heterosyllabic obstruent-obstruent sequences are still quite extensive by comparison to English, German, or Montreal French. The only obstruent type that is allowed in coda position is an unreleased lenis stop. When an underlying morpheme-final fricative or affricate or a fortis or aspirated stop occurs in a position where it cannot be resyllabified with a following vowel — i.e., when it is phrase-final or when the following morpheme begins with any consonant other than a glide — it surfaces as the homorganic lenis stop, as shown in the examples in Table 5. Thus, the only type of  $C_1V_1C_2V_2C_3V_3$  sequence where reanalysis alone would yield a phonotactically permissible coda is one where  $C_2$  is a lenis /p/, /t/, or /k/. However, this is exactly the context where a lenis stop would be voiced, making it unlikely that  $V_2$  would undergo devocalization. On the other hand, coda stops are also always unreleased. Therefore, the overlap on the oral tier must be extreme enough so that the closure for  $C_3$  can hide the aerodynamic consequence of releasing  $C_2$ , and such extreme overlap could conceivably result in reanalysis even when  $C_2$  is an underlying fortis or aspirated stop. However, as in Japanese, complete overlap in devocalization seems to be limited to the glottal tier, and syllable count is maintained in the face of devocalization.

Table 5. Coda neutralization in Korean.

/təs/ 'cover' + /posən/ 'cloth socks' → /tət.p'o.sən/ '(traditional cloth) slippers'
/pat <sup>h</sup> / 'aduki bean' + /komul/ 'flour' → /pat.k'o.mul/ 'aduki bean flour'
/tʃip/ 'house' + /k <sup>hi</sup> / 'key' → /tʃip.k <sup>hi</sup> / 'house key'
/natʃ/ 'to be low' + /ta/ non-past tense marker → /nat.t'a/ 'is low'

Adherence to otherwise unviolated phonotactic constraints is also an important factor in determining the syllable count for several of the English and German forms in Table 1. For example, reanalysis and loss of a syllable is precluded in [tʃi'ro] for *Toledo* and [ʃmp<sup>h</sup>anzə] for *Schimpanse*, where it would yield illegal onset clusters /tʃ/ and /ʃmp/. It is also precluded in [p<sup>h</sup>Δm] *pommel* and [laɪdŋ] *leiden*, where it would yield illegal coda clusters /ml/ and /dn/. That it is phonotactic constraints related to syllable position at work here is clear when we contrast the latter two forms to words such as *family* and *geladene*. In these originally three- or four-syllable forms, the same /ml/ and /dn/ sequences would result from reanalysis, but because the reduced vowels are in medial syllables in these cases, the consonants can be differentially assigned to the preceding and following syllables to make legal coda and onset. In these forms, reanalysis seems more possible, and indeed Zwicky (1972, p. 283) lists *family* as an example of "obligatory" application of his schwa-deleting rule.

However, the standard view of phonotactic constraints cannot explain all cases where reanalysis seems difficult. Reinterpreting *Arabic* and *element* as disyllabic /æ.ɪ.bɪk/ and /ɛl.mɪnt/ gives consonant sequences that are actually attested in the words *barber* and *Elmer*, yet Zwicky (1972) rules out even the application of the fast-speech reduction rule in these forms. He cites *Arabic* and *element*, along with *graciously*, *relevant*, and many others, as evidence that, unlike the rule of "Pre-stress Contraction" that gives rise to [vs'ɪnɪrɪ] for *vicinity*, "the deletion in Slur is not governed by any simple or obvious conditions on the 'pronounceability' of the result" (Zwicky 1972, p. 283). To account for these patterns, Zwicky posits a continuum of acceptability, with values based on the type of consonant following the reduced vowel, according the ranking:

[j] > [l] > [n] > [m] > [ŋ] > fricatives > stops  
 where ">" should be read as "is better than". That is, the output of the schwa-deletion rule is more acceptable (or the rule can be applied in a higher proportion of words) in which the following environment is an [j] than in words where an [l] follows, and so on. (Most of the forms where the rule is "obligatory" for Zwicky in fact are forms with following [j], and Gimson 1989, pp. 238-239, gives more examples with [j] and [l] than [n] or any other following consonant.)

My own intuitions accord with the notion that there is a continuum of acceptability governing both the likelihood of reanalysis and the degree of reduction that is tolerated, although I would emphasize also other things in the ranking. Zwicky himself lists two other things. Reanalysis seems considerably less likely if the following syllable has secondary stress (as in the verb *degenerate* [dɪdʒ'ɛnə.ɪ.ɛt] versus the adjective [dɪdʒ'ɛn.ɪ.ɛt]), or if the target vowel is in a closed syllable (as in *development* or *honestly*).

The first of these factors might be related to the ways in which prosodic structure is realized in the gestural score. For example, when the following consonant begins a stressed syllable, there is probably less overlap between it and the target vowel. Evidence for this lesser overlap can be seen in the greater coarticulation of an unstressed syllable with a preceding stressed syllable than with a following (e.g.

Fowler 1981) and the lack of a durational effect of voicing of the following consonant when the target vowel is reduced (Davis & Summers 1989).

I think the second factor has to do not so much with the mere fact of a cluster following the target vowel as it does with the phonological complexity or awkwardness of the resulting coda or onset, and I would also include the sonority difference between the flanking consonants as a possibly independent factor influencing the relative awkwardness. Thus, disyllabic *Arabic* (with a heterosyllabic sonorant-obstruent sequence /ʌ.b/) seems a more plausible phonological interpretation than disyllabic *element* (with its two sonorant consonants /l.m/), which in turn is far more plausible than disyllabic *honestly* (with its more complex coda cluster /nst.l/) or disyllabic *graciously* (with its tongue-twister sequence of postalveolar and alveolar fricatives /ʃ.sl/).

To be sure, more than complexity or awkwardness might be involved in *honestly*, since /t/ cannot be in the onset of a following syllable with /l/ and /nst/ is not an attested coda word internally. (It is arguably not even an attested coda word-finally, since coronals in forms such as in *minced* have been claimed to be affixed outside of the syllable core — see, e.g., Fujimura 1979.) However, none of the other cases cited here involves a clearly phonological restriction against the resulting consonant cluster: /ʃ/, /l/, and /ʃ/ are all attested post-vocalic strings in English, and /b/, /m/, and /sl/ are all attested onsets.

At the same time, it is not clear to me that these factors can be so easily distinguished from the considerations that preclude reanalysis in *Toledo* or *pommel*, or in the majority of Japanese and Korean cases cited earlier. The distinction between phonological and phonetic factors hinges crucially on a very rule-like conception of phonotactics (and of morpheme-structure constraints in general), a conception whereby a grammar either has a particular constraint or it does not. It might be fruitful to pursue alternative models of phonology in which traditional phonotactic constraints are only an extreme example of speakers' knowledge of the relative frequencies of sequences and structures in the lexicon (see Pisoni, Nusbaum, Luce, & Slowiaczek 1985; Pierrehumbert in press). That is, for an English speaker, the awkwardness of /l.m/ and the unacceptability of /nst.l/ word-medially might differ from the illegality of /tl/ word-initially not in kind but in degree: a low probability of occurrence versus zero or near-zero probability. In such a model the fact that reanalysis of a reduced vowel is impossible for [t'l'ido] *Toledo* and possibly for [ʃ'ɔnstli] *honestly* but merely quite unlikely for ['elmɪnt] *element* would not be two disparate facts to be represented by two different devices, but merely two points along the same scale.

Such a conception of phonotactic constraints might also offer a better understanding of the Japanese case, where a consonant preceding a devocalized vowel remains an onset even when reanalysis as a coda consonant would yield a phonotactically legal heterosyllabic consonant cluster. As Kubozono (1993) points out, the vast majority of syllables in Modern Japanese are open. Although closed syllables are phonotactically legal, they are comparatively rare. Reanalyzing the consonant in the second syllable of a form such as [ka.f.ʃj.tɕw] or [ʃo.k(i).kɔw] as a coda for a long syllable rather than as an onset for a short syllable may be unacceptable for a lesser degree of the same reason that reanalyzing the second consonant in [kafjta] as a coda is impossible. The analogous account should work for the Korean facts as well.

The distribution of "true syncope" in Montreal French, on the other hand, may remain difficult even within this framework. Phonotactic constraints on syllable structure do seem to affect the likelihood of devocalization. Cedergren and Simoncau (1985) list the indicated vowels in [obl(i)ʒe] *obligé* and [admin(i)stʁasjɔ] *administration* as cases where syncope is highly unlikely, and an adjacent consonant cluster is one of the strongest negative factors in their Varbrul analysis of

the frequency of occurrence of such completely elided vowels. If frequency of occurrence is related to likelihood of reanalysis, these facts would suggest that syncope is avoided where it would result in clusters such as [bɫɫ] and [nsts] that cannot be readily parsed in reanalysis. However, Cedergren and Simoneau also list

among their examples of likely syncope not only forms such as [pɫɫɔ] for *principaux*, [televzjɔ] for *télévision*, and [lynivekste] for *l'université*, where reanalysis would yield a perfectly legal consonant cluster, but also forms such as [pma] for *pis ma*, [vsave] for *vous savez*, and [tʒu<sup>u</sup> r] for *toujours*, where an impossible syllable onset would result. Presumably, these forms are not as readily reanalyzed as *principaux*. Yet their Varbrul analysis showed "true syncope" to be slightly more likely to occur in the initial syllable of a rhythm group than in other non-final positions. Since this is a position where such impossible clusters are likely to result (as in the examples just cited of *pis ma*, *vous savez*, and *toujours*), perhaps the likelihood of phonological reanalysis and the frequency of occurrence of such extreme devocalization may not be so closely related as we would expect from our explanation of the relative infrequency of devocalization in vowels before consonant clusters. Or perhaps there is some other factor that works to obscure the relationship between phonotactic acceptability and devocalization rates when comparing vowels in different positions in the rhythm group.

## 6. Some other factors

The preceding section examined phonotactic constraints on the shapes of syllables as a factor in determining whether a form with a substantially reduced vowel is susceptible to phonological reanalysis as having no vowel underlyingly — i.e., as having no syllable nucleus in that position. The discussion at several points hinted at other factors that might also be at work in influencing whether reanalysis will occur. There is no space to pursue the question further in this paper, but I can at least briefly mention three other factors that seem worth investigating in more detail.

The first factor is the existence of contrasting forms. This factor was already suggested in the observation that in Montreal French, *citation* [ʃ.ta.sjō] seems to be immune to reanalysis because of the potential homophony with *station* [sta.sjō]. Similarly in German, Stefanie Jannedy (personal communication) says that reanalyzing *geladene* as three syllables is easier for her than reanalyzing *Ebenen* as two, because there is no contrasting form \**geladne* but *Ebenen* contrasts with the related verb *ebnen* 'to level, to smooth'. On the other hand, the resulting homophony does not seem to preclude reanalysis for *Kannen*. Thus, existence of contrasting forms may be a relatively weak consideration at best.

The second factor is clearly stronger. It is the influence of other levels of linguistic structure — of prosodic constituents above the syllable. In German and English, the crucial relevance of one higher-level constituent is obvious already in the definition of the reduction effect. "Schwa-deletion" cannot occur unless there is a /ə/ — that is, it cannot occur in a syllable that is the head of stress foot. The different patterns of reanalysis seen in British and American English for words such as *laboratory* (three-syllable British [læb'ɔ<sup>ə</sup> tɔɪ] versus four-syllable American [l'æbɹət,ɔɪ]) highlight the importance of this prosodic constituent. However, the unlikelihood of complete reduction of a stressed vowel is not the only aspect of foot structure that is relevant in determining when reanalysis can occur. In northern German dialects, it seems that reanalysis is also more difficult for a pre-tonic syllable than for a post-tonic syllable. Thus, extreme vowel reduction and consequent assimilation and simplification of the resulting [gn] sequence is quite likely in casual-speech productions of *Wagen* (see [v'a:gn̩] → [v'a:gn̩] → [v'a:ŋ] in Table 2 above), but comparable reduction and simplification is impossible in *gemütlich* 'cozy' → \*[gm'ytlɪç] → \*[gŋ'ytlɪç] → \*[ŋ'ytlɪç]. The relevant fact here

seems to be the foot affiliation of the consonants; the two consonants in the resulting /gn/ sequence in *Wagen* belong to the same foot, whereas in *gemütlich* the /g/ does not belong to the foot that starts with the /m/. Here we can also contrast *Sekunde* /zɛkʷndə/ 'second' to *subtil* 'subtle'. In Stefanie Jannedy's Hamburg dialect *Sekunde* cannot lose its schwa to become \*[skʷndə], presumably because the [k] here is an onset consonant for the head of the stress foot, whereas the [p] in *subtil* is the coda for the pre-tonic syllable (as is clear from the devoicing of the original /b/ in the French loan source, which is still reflected in the orthography).

While the stress foot is quite obviously related to reduction and reanalysis in these two stress-timed languages, it is not the only potentially relevant prosodic unit cross-linguistically. As noted already in Section 5, the effects of position of the target syllable in the accentual phrase are particularly relevant for understanding why reanalysis does not occur for devocalized vowels in Korean. A post-lexical phonetic voicing of lenis stops that are medial to the accentual phrase removes the segmental environment for devocalization in just the position where reanalysis would yield a legal consonant coda. The homologous unit in Montreal French also strongly conditions the likelihood of vowel reduction. Again as noted already, devocalization does not occur at all in the last syllable of the rhythm group, and it is slightly more likely to occur in the first syllable than in other non-final syllables. It would be interesting to see whether initial syllables have the same advantage over medial syllables if penultimate syllables are excluded from the comparison. Levac et al. (1993) show that in Montreal French, the lengthening seen cross-dialectally in the last syllable of the rhythm group also stretches out the immediately preceding syllable, albeit to a somewhat lesser degree. That is, the lengthening at the end of this prosodic constituent in Montreal French is a true edge effect, comparable to the final lengthening in intonation phrases in English (e.g. Edwards, Beckman & Fletcher 1991), and rather unlike the lengthening in the European dialects of French, where the longer final syllable seems instead to be the culmination of a rhythmic alternation involving a shorter penultimate syllable and a possibly longer phrase-initial syllable bearing an *accent secondaire* (see, e.g., Fletcher 1991).

This difference between Montreal French and Parisian French seems related to the third factor: the sociolinguistic significance of the reduction. If forms with extremely reduced vowels are associated with one particular dialect and not another, then all of the sociolinguistic considerations of the relationship between the two dialects come into play, and we might expect reanalysis to be generalized, in either of two ways. On the one hand, if extreme reduction is associated with a more standard dialect, we might expect the reduction to become phonologized as the standard spreads. That is, if the phenomenon is salient as a mark of a more prestigious dialect to speakers of a nonstandard dialect, they might abstract a categorical rule for deriving forms in the standard dialect from the homologous native forms. Alternatively, if extreme reduction is associated with a stigmatized regional or class dialect, then the phenomenon might again become phonologized as a categorical mark of familiarity and group solidarity. Either pattern leads to the same result: vowel reduction (or any other connected speech process) should become phonetically less variable when it acquires sociolinguistic significance in relationship to another community of speakers (cf. Nolan & Kerswill 1990).

Something like this hypothesis of phonologization seems to be assumed in Cedergren and Simoneau's (1985) discussion of their initial Varbrul analysis of overall levels of devocalization in their Montreal French speakers. The analysis included such factors as the age, education, and sex of the speaker, and the speaker's level of participation in the "linguistic market" — i.e., whether the speaker's work required that he or she interact often with speakers using a less regionally marked variety of French (Sankoff & Laberge 1978). The women overall had generally lower rates for each of the three grades of reduction than did

the men. Their productions also showed a closer relationship with degree of participation in the linguistic market than to age, and a much more complex pattern of phonological conditioning. Cedergren and Simoneau interpret the comparatively simpler phonological conditioning for the men (and the stronger relationship to age)

as indicating "un phénomène en début d'expansion", whereas "pour les femmes le processus est plus avancé dans le temps et servirait plutôt de différenciateur social." (Cedergren & Simoneau 1985, p. 76). If a more complex phonological conditioning means a phonetically more variable phenomenon, then this suggests that the phenomenon is more phonologized among younger male speakers for whom it is spreading as a mark of solidarity in opposition to "standard" French.

It would be interesting to see also whether solidarity and prestige can account for some of the patterns of variability in vowel reduction in German. As opposed to the more northern varieties reported in Kohler (1990) and Jannedy (1993), southern German dialects show much more wide-spread reduction in pre-tonic syllables. In replicating her earlier experiment, this time using Schwäbisch speakers, Stefanie Jannedy has got recordings of productions such as [kʃt'ɔ:ə] for *gestohlen* 'stole', [pʃ'aft] for *beschafft* 'procure', and [kʃ'ʋesə] for *gefressen* 'eaten', even at normal rate.<sup>2</sup> Has the pre-tonic reduction been reanalyzed as a phonological mark of solidarity for these southern speakers? Conversely, do northern speakers deliberately retain such pre-tonic syllables in order not to sound like southern speakers? Has devoicing become so ubiquitous in Tokyo Japanese because it is a mark of the prestigious standard? The fact that the national broadcasting corporation has twice updated its dictionary prescribing where to devoice (NHK 1985) would suggest some such sociolinguistic significance. Conversely, my impression is that Kyoto and Osaka speakers very deliberately voice the vowels in the polite adjectival and verbal endings *-desu* and *-masu* when excluding speakers of the standard language.

The five languages briefly reviewed in this paper seem fruitful ground for investigating the many questions that become thinkable when we adopt an account of vowel reduction that represents it originally as the aerodynamic consequence of overlap in the gestural score, with potential phonological reanalysis for forms in which extreme overlap is a habitual pattern for some prosodic context or some group of speakers.

### References

- Beckman, M.E., de Jong, K., Jun, S.-A., & Lee, S. (1992). The interaction of coarticulation and prosody in sound change. *Language and Speech* 35, 45-58.
- Beckman, M., & Shoji, A. (1984) Spectral and perceptual evidence for CV coarticulation in devoiced /si/ and /syu/ in Japanese. *Phonetica* 41, 61-71.
- Bloch, B. (1950) Studies in colloquial Japanese IV: Phonemics. *Language* 26, 86-125.
- Browman, C. P., & Goldstein, L. (1990a) Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman, eds.,

<sup>2</sup>Note that the post-tonic /ə/ can remain unreduced in forms such as *gestohlen* and *gefressen* even at rates where the pre-tonic readily reduces to nothing. This is the opposite pattern from the analogous forms in northern German, where the pre-tonic syllable in, say, *geleiten* cannot be reduced unless the post-tonic is reduced at least to the same extent. Presumably the difference is related to the dialectal deletion of the final /n/ in *gestohlen* and *gefressen*, since in northern German forms such as *gerade* 'just', where there is no coda consonant in the post-tonic syllable, the facts are similar to those in southern German.

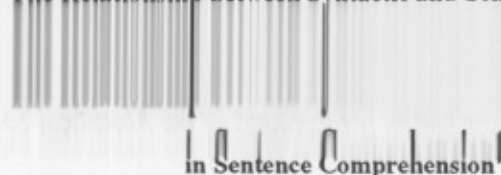


- Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pp. 341-376. Cambridge: Cambridge University Press.
- Browman, C. P., & Goldstein, L. (1990b) Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics* 18, 299-320.
- Cedergren, H. J., & Simoneau, L. (1985) La chute des voyelles hautes en français de Montréal: "As-tu entendu la belle syncope?" In M. Lemieux & H. J. Cedergren (eds.) *Les tendances dynamiques du français parlé à Montréal*, pp. 57-144. Montreal: Office de la langue française.
- Davis, S., & Summers, V. W. (1989) Vowel length and closure duration in word-medial VC sequences. *Journal of the Acoustical Society of America* 85, S28.
- Delattre, P. (1951) Le jeu de l'E instable intérieure en Français. *French Review* 24, 341-351.
- Edwards, J., Beckman, M. E., & Fletcher, J. (1991) Articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America* 89, 369-382.
- Fletcher, J. (1991) Rhythm and final lengthening in French. *Journal of Phonetics* 19, 193-212.
- Fokes, J., & Bond, Z. S. (1993) The elusive/illusory syllable. *Phonetica* 50, 102-123.
- Fowler, C. A. (1981) A relationship between coarticulation and compensatory shortening. *Phonetica* 38, 35-50.
- Fujimura, O. (1979) An analysis of English syllables as cores and affixes. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 32, 471-476.
- Gimson, A. C. (1989) *An Introduction to the Pronunciation of English* (4th ed., rev. by S. Ramsaran). London: Edward Arnold.
- Hall, T. A. (1992) *Syllable Structure and Syllable-related Processes in German*. Linguistische Arbeiten 276, Tübingen: Max Niemeyer Verlag.
- Hirose, H., Lee, C. Y., & Ushijima, T. (1974) Laryngeal control in Korean stop production. *Journal of Phonetics* 2, 145-152.
- Itô, J., & Mester, A. (1993) Stem and word in Sino-Japanese: a case study in syllable optimization and alignment. Paper presented at the Dokkyo International Forum on Speech Recognition and Phonology, Dokkyo University, Soka City, Saitama Prefecture, 18-19 December 1993.
- Jannedy, S. (1993) Rate effects on German unstressed syllables. Colloquium paper, Department of Linguistics, Ohio State University.
- Jun, S.-A. (1990a) The domains of laryngeal feature lenition effects in Chonnam Korean. *Journal of the Acoustical Society of America* 87, S123.
- Jun, S.-A. (1990b) The prosodic structure of Korean — in terms of voicing. In E.-J. Baek (ed.) *Proceedings of the Seventh International Conference on Korean Linguistics*. Toronto: University of Toronto Press.
- Jun, S.-A. (1993) *The Phonetics and Phonology of Korean Prosody*. Ph.D. dissertation, Department of Linguistics, Ohio State University.
- Jun, S.-A., & Beckman, M. E. (1993) A gestural-overlap analysis of vowel devoicing in Japanese and Korean. Paper presented at the 1993 Annual Meeting of the Linguistic Society of America, 7-10 January 1993, Los Angeles, CA, USA.
- Kagaya, R. (1974) A fiberoptic and acoustic study of the Korean stops, affricates and fricatives. *Journal of Phonetics* 2, 161-181.
- Kohler, K. (1990) Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In W. J. Hardcastle & A. Marchal, eds., *Speech Production and Speech Modelling*, pp. 69-92. Amsterdam: Kluwer.

- Kondo, M. (1993) The effect of blocking factors and constraints on consecutive vowel devoicing in Standard Japanese. Poster presented at the Fourth Conference on Laboratory Phonology, Oxford, 11-14 August 1993.
- Kubono, H. (1989) The mora and syllable structure in Japanese: evidence from speech errors. *Language and Speech* 32, 249-278.
- Kubozono, H. (1993). Perceptual evidence for the mora in Japanese. Paper presented at the Fourth Conference on Laboratory Phonology, Oxford, 11-14 August 1993.
- Kubozono, H. (forthcoming) The syllable in Japanese. Submitted to *Language*.
- Kuehn, D. P., & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics* 4, 303-320.
- Levac, L., Cedergren, H. J., & Perreault, H. (1993) Phonetic evidence of narrow and wide temporal scope for prosodic constituents in French. In D. House & P. Touati (eds.) *Proceedings of an ESCA Workshop on Prosody [Working Papers, Department of Linguistics and Phonetics, Lund, No. 41]*, pp. 54-57.
- McCawley, J. D. (1968) *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton.
- Maekawa, K. (1990) Hatsuwa sokudo ni yoru yuusei kukan no hendoo [Effects of speaking rate on the voicing variation in Japanese]. Technical Report SP89-148, Densi Zyoohoo Tuusin Gakkai.
- Manuel, S. Y., Shattuck-Hufnagel, S., Huffman, M., Stevens, K. N., Carlson, R., & Hunnicutt, S. (1992) Studies of vowel and consonant reduction. *Proceedings of the 1992 International Conference on Spoken Language Processing*, vol. 2, pp. 943-946.
- Nihon Hoosoo Kyookai (1985) *Nihongo hatuon akusento jiten [Dictionary of Japanese Pronunciation and Accent Patterns]*, rev. ed. Tokyo: Nihon Hoosoo Shuppan Kyookai.
- Nolan, F., & Kerswill, P. E. (1990) The description of connected speech processes. In S. Ramsaran (ed.) *Studies in the Pronunciation of English: a Commemorative Volume in Honour of A. C. Gimson*, pp.295-316. Padstow, Cornwall: T. J. Press, Ltd.
- Ohala, J. J. (1974) Experimental historical phonology. In J. M. Anderson & C. Jones (eds.) *Historical Linguistics II: Theory and Description in Phonology*, pp. 353-389. Amsterdam: North-Holland.
- Ohala, J. J. (1981) The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (eds.) *Papers from the Parasession on Language and Behavior*, pp. 178-203. Chicago: Chicago Linguistic Society.
- Ohso, M. (1973) A phonological study of some English loan words in Japanese. *Ohio State University Working Papers in Linguistics* 14, 1-26.
- Pierrehumbert, J. (in press) Syllable structure and word structure: a study of triconsonantal clusters in English. In P. A. Keating (ed.) *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology 3*. Cambridge: Cambridge University Press.
- Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985) Speech perception, word recognition and the structure of the lexicon. *Speech Communication* 4, 75-95.
- Price, P. J. (1980) Sonority and syllabicity. *Phonetica* 37, 327-343.
- Sankoff, D. & Laberge, S. (1978) The linguistic market and the statistical explanation of variability. In D. Sankoff (ed.) *Linguistic Variation: Models and Methods*, pp. 227-238. New York: Academic Press.
- Shibata, T. (1955) Museika [Devoicing]. In *Kokugogaku jiten [Dictionary of Japanese Language Studies]*, p. 899. Tokyo: Kokugogaku Gakkai.
- Silva, D. J. (1992) *The Phonetics and Phonology of Stop Lenition in Korean*, Ph. D. dissertation, Department of Linguistics, Cornell University.

- Strauss, S. L. (1982) *Lexicalist Phonology of English and German*. Dordrecht: Foris.
- Takeda, K., & Kuwabara, H. (1987) Boin museika no yoin bunseki to yosoku syuhoo no kentoo [Analysis and prediction of devocalizing phenomena]. *Nihon Onkyoo Gakkai Kooen Ronbunshu [Proceedings of the Acoustical Society of Japan]* 3-3-8, 105-106.
- Verluyten, S. P., ed. (1988) *La phonologie du schwa français*. *Linguisticae Investigationes Supplement*. Amsterdam: John Benjamins.
- Zwicky, A. M. (1972) Note on a phonological hierarchy in English. In R. Stockwell & R. Macauley, eds., *Linguistic Change and Generative Phonology*, pp. 275-301. Bloomington, IN: Indiana University Press.

## The Relationship between Syntactic and Semantic Processes



Julie E. Boland

jboland@ling.ohio-state.edu

**Abstract:** Two experiments investigate how lexically ambiguous input is handled by the sentence processing system and shed light on the relationship between syntactic and semantic processing. Sentence contexts containing ambiguous verbs (e.g., *Which salad/baseball did Janet toss...* probe-word: *Bill*) are used to investigate how subcategorization and thematic role information is used by the sentence processing system. The results are consistent with a model in which the syntactic processing system uses subcategorization information to compute all "legal" structures in parallel, without consideration of semantic information from the context. Meanwhile, the semantic processing system uses contextual information to pursue the single most likely semantic analysis. The resulting syntactic and semantic representations are checked against each other, and inconsistent analyses discarded.

Models of sentence understanding often decompose the task into syntactic and semantic processes. The degree to which syntactic and semantic processes are independent, and the relationship between them, are matters of great debate. Tangled up in this debate is the role of combinatory lexical information in the initial stages of sentence understanding. This is because a verb's "subcategorization frames" and "thematic structure" are part of the lexical knowledge made available when a verb is recognized. A subcategorization frame is a syntactic representation of a verb's arguments.<sup>1</sup> A thematic structure is a representation of the "thematic roles" assigned by a verb, thematic roles being generalized characterizations of an argument's mode of participation in the event described by a sentence.

Views regarding how combinatory lexical information is used by the sentence processing system vary widely. The "Garden Path" model maintains that there is a

---

\*Experiment 1 was completed at the University of Rochester as part of the author's doctoral dissertation and was supported by grant HD-27206 (to Michael Tanenhaus). Experiment 2 was conducted at the Ohio State University and was supported by a University Seed Grant. The author gratefully thanks Michael Tanenhaus, Chris Barker, Kim Darnell, and three anonymous reviewers for many helpful comments on early versions of this manuscript.

<sup>1</sup>I am distinguishing arguments from adjuncts, which are always optional, and not subcategorized.

syntactic "module" which operates independently of semantic knowledge, excluding contextual and most lexical information from initial syntactic processing (e.g., Frazier, 1987, 1989, Mitchell, 1989). The initial parse is influenced only by major syntactic category, phrase structure rules, and a heuristic specifying that the simplest structure will always be constructed first. Because relevant information is initially ignored, the parser will make frequent mistakes, or "wander down the garden path". The opposing view is that the syntactic and semantic systems are completely integrated, with the extreme version making no distinction between syntactic and semantic processes (Bates & MacWhinney, 1982; Waltz & Pollack, 1985). Marslen-Wilson and Tyler (1987) take a more moderate position, in which theoretically distinct syntactic and semantic systems work interactively. A similar position has been taken by Tanenhaus and colleagues (e.g., Tanenhaus et al., 1989). They argue that subcategorization and thematic information are both used in the early stages of sentence comprehension, leading to the prediction that garden paths should occur only when lexical knowledge is ambiguous.

The current paper presents a processing model that requires still more moderation of the Interactive position. In the proposed model, the syntactic processing system functions independently of semantic information from the context. However, unlike the Garden Path model, the syntactic processing system uses subcategorization information to compute all "legal" structures in parallel. Meanwhile, the semantic processing system uses information from the output of the syntactic system and a variety of other sources (the lexicon, the discourse model, etc) to pursue the most likely semantic analysis.<sup>2</sup> The proposed model might be termed a "Concurrent" model, because the syntactic and semantic representations are computed simultaneously.

The Concurrent model is appealing on theoretical grounds for several reasons. First, because context is ignored by the syntactic system, syntactic representations can be built reflexively using phrase structure rules (or the equivalent) and syntactic category and subcategory information from the lexicon. When multiple structures are consistent with lexical information, multiple structures are automatically built. Little cost should result from building multiple representations because the structures are output automatically, in parallel. In contrast, semantic representations are heavily context-dependent, and their construction is a resource-intensive process. However, precisely *because* they are context-dependent, the context guides the construction of the single most likely semantic representation. If the initial semantic representation proves inconsistent with new information (such as the following words), the alternative syntactic representations may provide a mechanism for recovery.

The Concurrent model is similar to a class of models that can be described as "Parallel Syntax" models. A growing number of researchers are finding evidence for parallel postulation of syntactic structures, possibly using subcategorization

---

<sup>2</sup>In future versions of this model, it may be possible to specify how various properties of the context bear on the decision of which semantic analysis to pursue, but currently, only "thematic" properties are considered. When the verb is recognized, the set of possible thematic frames are compared with the current syntactic representation and the semantic properties of the current arguments. The best-fitting thematic frame is selected as the basis for the semantic analysis.

information, just like the Concurrent model (e.g., Crain & Steedman, 1985; Gorrell, 1989 & 1991; McElree, 1993). However, most advocates of the parallel models seem to assume, as does Frazier (1987), that the syntactic analysis must precede the semantic analysis, and some of them make specific claims about the

order in which the "parallel" syntactic representations undergo semantic processing. Like the Garden Path model, these models are essentially modular. Under the Concurrent model, it is possible for a semantic representation to be produced before the parser has settled on a single syntactic structure. Such an outcome is likely when subcategorization information permits computation of multiple syntactic representations, but contextual information strongly biases a single interpretation. While such an outcome *may* be possible with a Parallel Syntax model, it is clearly not possible with the Garden Path model. The current experiments do not allow comparisons that would clearly distinguish between the Concurrent model and some Parallel Syntax models. Therefore, the contrast will be between the Concurrent model, the Garden Path model, and Interactive models in which semantic information influences the initial parse.

In order to distinguish between the Concurrent model, the Garden Path model, and Interactive models, one must have the methodological tools to distinguish between syntactic and semantic representations in a way that is not confounded with theoretical assumptions. For example, it has sometimes been implied that first pass eye-movements are indicative of first-pass (i.e., syntactic) parsing and regressive/second pass eye-movements are indicative of later (i.e., semantic) interpretive processes (e.g., Ferreira & Henderson, 1990). While this is perfectly plausible, it is linked to the theoretical claim that syntactic processing precedes semantic processing. There are no pre-theoretical grounds for identifying first pass eye movements with purely syntactic processes. In contrast, recent work measuring Event-Related Brain Potentials (ERPs) has suggested that distinct components of the waveform can be linked to syntactic and semantic anomalies (Hagoort et al., 1993; Neville et al., 1991; Osterhout & Holcomb, 1992). Further exploration of this paradigm may yield a paradigm in which semantic and syntactic processes could be distinguished experimentally.

The current studies use the cross-modal Integration Paradigm introduced in Boland (1991, see also Boland, 1993)<sup>3</sup> as a partial solution to the methodological problem of distinguishing between syntactic and semantic representations. Auditory presentation of a sentence or sentence fragment is followed by a visual target that may or may not be a good continuation of the sentence. The assumption is that as we hear or read a sentence, we immediately begin constructing syntactic and semantic representations of it. Responses to the target word will be faster when the target is consistent with the relevant representation than when it is not. For reasons that are discussed at length in Boland (1991), the ability to integrate the target into the syntactic representation is most relevant when the task is naming, but both the syntactic and semantic representations play a role when the task is lexical decision. Note that naming and lexical decision are

---

<sup>3</sup>Experiment 1 of the current paper is Experiment 3 in Boland (1991). This experiment was also briefly summarized as Experiment 2 in Boland (1993).

used here to investigate **post-access** effects -- namely, the ease with which the target can be integrated into the sentence context.

The Integration Paradigm contrasts naming and lexical decision tasks based on the evidence that naming is most sensitive to syntactic representations and lexical decision is sensitive to both syntactic and semantic representations. However, this distinction is probably not absolute. For example, recent studies in my own laboratory using only visual representation have not obtained clear patterns of results. The conditions under which this task difference can be obtained must be better understood in order for this paradigm to be fully convincing. The current experiments handle this potential difficulty by including control conditions designed to determine which level(s) of representation are being tapped.

### Experiment 1

Experiment 1 was designed to investigate how verb-specific syntactic and thematic knowledge become available to the sentence processing system and how this lexical knowledge influences on-line construction of syntactic and semantic representations of sentences. The focus is on verbs such as *toss*, that have multiple argument structures (i.e., *toss the ball to the child vs toss the salad*). If the entire lexical entry is made available at once, all the syntactic and semantic knowledge associated with each of a verb's senses and argument structures would presumably become available. How might the sentence processing system sort out this information and make use of it?

All models of sentence processing assume that at least some aspects of lexical knowledge are used in the initial stages of sentence processing. For example, the Garden Path model assumes that major grammatical category is used (along with phrase structure rules) to assign each incoming word to the phrase structure tree. Interactive models maintain that subcategory and thematic role information are used for initial parsing decisions. In either case, the output of the word recognition system provides crucial input to the sentence processing system. However, the output of the word recognition is often ambiguous and it is not clear how the sentence processing system would handle lexically ambiguous input.

Consider the way lexical access is believed to occur. There is general agreement (Forster, 1979; Marslen-Wilson, 1987) that lexical items are accessed in a bottom-up fashion when linguistic input is perceived; contextual information cannot access lexical items independently, nor does context restrict lexical items from being accessed by the input. Thus, when physical input is ambiguous, multiple lexical forms are accessed, although not always simultaneously (see for example, Van Petten & Kutas, 1987). Context is then used to select the appropriate candidate.<sup>4</sup> However, the most common current views of sentence processing maintain that the parser constructs syntactic representations in serial

---

<sup>4</sup>However, if one sense of an ambiguous word is accessed earlier than the other senses and it can be quickly integrated into the context, the alternative senses may not be accessed (Rayner & Morris, 1991). For example, Tabossi (1988) found that a subordinate meaning will not be accessed if the appropriate semantic features of the more frequent meaning are primed by the context.

(e.g., Frazier, 1987). A problem arises because a large number of English words are ambiguous at some level. Even if the parser only makes use of syntactic category, what can it do with noun/verb ambiguities like *ring* or adjective/noun ambiguities like *green*? And if the parser makes use of subcategory information in

addition to major syntactic category, the ambiguity is multiplied because many verbs allow multiple subcategorization frames. Is the appropriate subcategorization and thematic frame selected in the same way that the appropriate meaning of an ambiguous word is selected?

Experiment 1A uses cross-modal naming and was designed to determine whether multiple subcategorization frames are made available when verbs with multiple senses are recognized. ("Sense" is loosely defined here as a difference in the type of event denoted by the verb.) Experiment 1B uses cross-modal lexical decision to ask the same question about thematic frames. The critical sentences contain verbs with senses that have different numbers of arguments associated with them. For example, the sense of *toss* associated with salads has just two arguments, a subject and an object. In contrast the "throw" sense can have three arguments: subject, direct object, and indirect object. Thus, (1a) is unacceptable<sup>5</sup>, but (1b) is fine. Remember, according to the lexical access literature, semantic associates of both senses would be facilitated at the offset of *toss* in both (1a) and (1b). Several hundred milliseconds later, only the contextually appropriate probe would be facilitated. This is because words are accessed in bottom-up fashion -- then context is used to select the most likely sense.

1a) ?\*Which salad did Jenny toss Bill?

b) Which baseball did Jenny toss Bill?

Presumably the entire lexical entry for each sense of an ambiguous word is activated, not merely multiple meanings. It ought to be possible to design an experiment that tests for multiple activation of lexical argument structures that is exactly analogous to the semantic priming experiments that test for multiple activation of meanings. If multiple argument structures are made available, then multiple syntactic and thematic representations might initially be formed at words like *toss*. Thus, integrating the target, *BILL*, into the contextually inappropriate representation ought to be equivalent to integrating the target into the contextually appropriate representation, but only if the target is presented during the window of time when both representations are available. Remember, integration effects in naming are likely to reflect syntactic integration, whereas integration effects in lexical decision may reflect both syntactic and semantic integration.

Two control conditions using "unambiguous" verbs are necessary to ascertain whether the task is tapping syntactic representations, semantic representations, or both. The first, illustrated in (2a), used simple transitive verbs that only subcategorize for two arguments: a subject and a direct object. This condition

<sup>5</sup>Note that (1a) is somewhat implausible, but acceptable if the "throw" sense is adopted. Some speakers may also find (1a) acceptable as the short form of *Which salad did Jenny toss for Bill?* There was some variation among the materials regarding how strongly the indirect object biased one meaning over the other and whether or not a benefactive reading was possible.



served as the baseline at which neither syntactic nor thematic integration is possible. The other control condition (2b) distinguishes between semantic integration and syntactic integration by using non-alternating datives. These verbs have three thematic roles and subcategorize for a prepositional indirect object, so that *SAM* is thematically congruent, but syntactically incongruent. If naming latencies for the two control conditions are equivalent and longer than naming latencies for the "baseball" condition, it will verify that naming is only sensitive to syntactic integration, and not thematic integration. Lexical decision latencies will be shorter for the thematically congruent condition if the task is sensitive to thematic integration.

- 2a) *Which necklace did Nancy touch.. SAM*
- b) *Which necklace did Nancy describe.. SAM*

Note that *toss* is an alternating dative that can take a noun phrase indirect object, so the ambiguous three argument condition (illustrated by (1b)) should allow both syntactic and thematic integration under any account. If it is significantly faster than the unambiguous two argument condition (2a), that will provide evidence for an integration effect in either task. In addition, it will provide a standard against which to compare the ambiguous two argument and the unambiguous three argument conditions. These are the two most interesting conditions. If argument structure information follows the same pattern of "activation" as semantic associates, then multiple sets of argument structure information should initially be available. This would be reflected by response times in the ambiguous two argument condition equivalent to those in the ambiguous three argument condition.

Note that it is possible to continue the "incongruent" conditions in such a way that the context+target is a legal string. Some examples are given in (3), below. However, in each of these cases, the target word is the direct object and the wh-phrase is part of an adjunct phrase. A number of researchers have used a variety of paradigms to examine how fronted wh-phrases are analyzed in sentences like those used in this experiment. All the evidence demonstrates that the wh-phrases are assigned the direct object role when a transitive verb is encountered (Clifton et al., 1984; Frazier & Clifton, 1989; Garnsey et al., 1989; Kurtzman, 1989). It is certainly possible that subjects would construct just such a structure when they encounter the target. But doing so would require some reanalysis (of the wh-phrase) and thus response times should be longer in these conditions compared to "congruent" conditions in which no reanalysis is necessary.

- 3a) *Which salad did Jenny toss Bill the croutons for?*
- b) *Which necklace did Nancy touch Sam with?*
- c) *Which necklace did Nancy describe Sam wearing?*

The choice of proper names as targets has two consequences. First it insures that all targets are equally unpredictable so that subjects cannot generate potential targets from the context. Second, it required that the traditional lexical decision

instructions be modified slightly because the targets were names rather than words. I have assumed that common first names are represented in the lexicon somewhat like nouns -- thus, it is possible to access the lexical item along with the semantic features of human, animate, male, etc.

#### Experiment 1A

The Concurrent model differs from the Garden Path model and serial Interactive models with regards to the number of syntactic representations that are constructed. The Concurrent model predicts that all subcategorized structures are pursued in parallel. In contrast, the Garden Path model maintains that only the syntactically simplest structure is initially constructed -- without regards to subcategory information. At the other extreme, serial Interactive models maintain that only the most contextually plausible syntactic representation will be constructed.

Experiment 1A tests one hypothesis of the Concurrent model, namely, that syntactic representations corresponding to each subcategorization structure are initially constructed when verbs with multiple subcategorization frames are encountered. This predicts that naming times in the two ambiguous verb conditions will be equivalent, and faster than the unambiguous verb conditions, which are both syntactically incongruent. In contrast, serial Interactive models predict that only the contextually appropriate structure will be constructed, so responses should be faster in the ambiguous three argument condition than the ambiguous two argument condition. These predictions are summarized in (4a) and (4b), respectively.

The predictions of the Garden Path model are less clear. Subcategory information, specifying that a third argument is possible, would not be available at the point when the target is presented. Thus, there might be no representation into which the target could be easily integrated, causing the ambiguous conditions to be equivalent to the unambiguous conditions. Alternatively, if subjects took the target to be part of the sentence, the simplest attachment uses the double object structure associated with (5b), below. Because subcategory information is not available, this structure would be used for both ambiguous and unambiguous verbs -- again predicting no differences across the four conditions as shown in (4c). To insure that the task is tapping the earliest point in processing, when (according to the Garden Path model) subcategory information is not available, the visual target was presented just before the auditory offset of the verb. Note, however, that if subcategory information became available in time to influence the naming response, one might find exactly the pattern predicted by the Concurrent model.

- 4a) Predictions of the Concurrent model:  $3A = 2A < 3U = 2U$
- b) Predictions of serial Interactive models:  $3A < 2A = 3U = 2U$
- c) Predictions of the Garden Path model:  $3A = 2A = 3U = 2U$

The unambiguous verb conditions also test the methodological hypothesis that naming is insensitive to semantic congruity. The prediction is that naming times for the two unambiguous conditions will be equivalent, with no advantage for the

three argument condition. Furthermore, a difference between the ambiguous and unambiguous argument conditions will provide evidence that subcategorization information is available early. The Concurrent model and the Interactive model both require that subcategory information be available early, while the Garden Path model requires that there be some time-point after word recognition when subcategorization information is not yet available.

#### *Method.*

**Subjects.** Forty undergraduates at the University of Rochester completed the experiment in partial fulfillment of course requirements or for a nominal fee. All were native speakers of English.

**Materials.** The Ambiguous conditions use ten alternating dative verbs that were judged to have another sense in which they were two argument transitives. For each verb, two versions of a sentence fragment were constructed. The two versions were identical except for the fronted, wh-phrase that was a filler for the direct object gap. In each pair, one of the wh-phrases strongly biased the two argument meaning, and the other strongly biased the three argument meaning. Ten additional sentence fragments were constructed, using the same structure, for the control conditions. Each of these fragments was also made into two versions, which were identical except for the verb. The verb was an unambiguous two argument verb in one version and a non-alternating dative (three argument verb) in the other. A sample set of experimental sentences is shown in (5). The full set is listed in the Appendix. Sentence completion norms were collected on all the experimental contexts to insure that the contexts biased the ambiguous verbs as expected and to insure that my judgments regarding verb subcategorization were appropriate.

- 5a) Ambiguous 2-Argument: *Which salad did Jenny toss.. BILL*
- 5b) Ambiguous 3-Argument: *Which baseball did Jenny toss.. BILL*
- 5c) Unambiguous 2-Argument: *Which necklace did Nancy touch.. SAM*
- 5d) Unambiguous 3-Argument: *Which necklace did Nancy describe.. SAM*

Ten sentence fragments with ambiguous verbs and ten with unambiguous verbs appeared on each of two lists and the two and three argument conditions of each were rotated between lists. In addition, 58 distractor fragments were constructed, about 30% of which were obviously cut off in mid-sentence. Targets for all trials were common first names, 2-5 letters in length. Targets for the critical trials were all single syllable names, 3-4 letters in length. Naming norms were collected on the targets without contexts. Targets in the ambiguous verb group averaged 393 milliseconds and targets in the unambiguous verb group averaged 390 milliseconds (N=8). There was no difference between the two groups of targets.

The sentence fragments were read into a tape recorder. An attempt was made to read all the critical fragments, as well as those distractors which ended mid-sentence, with neutral (as opposed to sentence-final) intonation. The materials were then digitized using the MacRecorder system. Five millisecond tones at 1000

hz were placed on the non-voice channel approximately 150 milliseconds before the offset of the verb or at the onset of the final consonant. In all cases, the tone occurred after the subjective recognition point of the word.

Procedure. Subjects wore headphones and were seated in front of a computer

screen, response box, and microphone. Contexts were presented through the headphones to both ears, and the target names were centered in all capital letters on the computer screen. Subjects responded by pronouncing the name into the microphone, which was connected to a noise-sensitive switch on the response box. Reaction times were collected from the time when the target came onto the screen until the noise-sensitive switch was triggered. If no response was registered within 2 seconds, the response was considered a time-out. In this paradigm, time-outs usually reflect mechanical trigger-failures rather than slow responses. Yes/No comprehension questions were presented visually after 25% of the trials to insure that subjects attended to the auditory contexts. In no case was it necessary to integrate the target with the context to answer the comprehension question. Subjects completed 10 practice trials, half of which had comprehension questions, before going on to the 78 experimental trials.

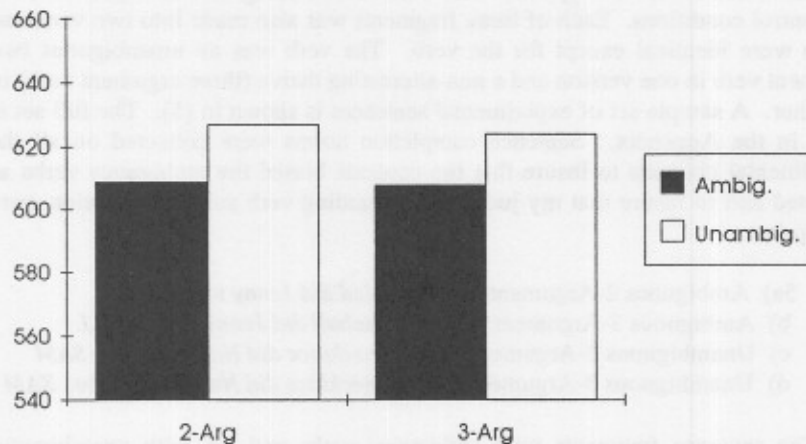


Figure 1. The mean naming latencies for each condition of Experiment 1A are given in milliseconds.

#### Results.

For each of the critical trials, the onset latency of the naming response was recorded. Time-outs accounted for less than 3% of the data. Within each condition, mean response times were computed by subject and by item. Responses were considered outliers if they were more than 2.5 standard deviations from a subjects mean response time. Outliers were replaced with the boundary value. About 4% of the data were replaced in this way. Mean response times are displayed in Figure 1. Subject and item means were each subjected to a 2(list) x 2(verb type) x 2(argument number) Analysis of Variance (ANOVA). There was a

main effect of verb type by subjects and by items [ $F(1,38)=9.58$ ,  $F(1,16)=8.68$ ,  $p < .01$ ], with responses to the unambiguous conditions slower than the ambiguous conditions. Importantly, there was no effect of argument number, which would have reflected semantic integration [ $F_s < 1.0$ ], and no interaction between verb type and argument number [ $F_s < .10$ ]. In a planned comparison of the three argument conditions, the ambiguous condition was faster than the unambiguous condition [ $F(1,38)=4.80$ ,  $F(1,16)=7.84$ ,  $p < .05$ ]. The difference between the two argument conditions was marginally significant [ $F(1,38)=3.90$ ,  $F(1,16)=3.63$ ,  $p < .10$ ]. The two ambiguous conditions did not differ from one another, nor did the two unambiguous conditions.

### *Discussion.*

Both theoretical and methodological predictions were supported. There was no difference between the two unambiguous conditions, and the ambiguous three argument condition was faster than the unambiguous three argument condition. This provides evidence that the task is insensitive to semantic congruity, but sensitive to subcategorization information. The Concurrent model's prediction that both subcategorization frames would be constructed was also supported. Naming times for the ambiguous verb conditions were fast compared to the unambiguous verb conditions. This suggests that the two argument and three argument conditions were both syntactically congruent -- and the two argument condition could not have been syntactically congruent unless the inappropriate subcategorization frame was available. Thus, the pattern of results is inconsistent with the predictions of the Interactive model. Furthermore, the results are not consistent with the Garden Path model unless it is the case that subcategory information becomes available in time to influence the naming response. What makes this unlikely, is the early point at which the target was presented. If subcategorization information is not available in time for the initial parse, it is not clear how it could be available soon enough to influence the naming response. Further evidence against the Garden Path model is provided by Experiment 1B and Experiment 2.

### Experiment 1B

The evidence from Experiment 1A, using naming, suggests that the parser constructed a structural representation corresponding to each subcategorization frame of the ambiguous verbs. Experiment 1B, which uses lexical decision, provides an opportunity to replicate that effect (because the lexical decision task is sensitive to syntactic congruity) as well as to test the hypothesis that only a single thematic frame is pursued (because the lexical decision task is also sensitive to semantic congruity). I am assuming that all thematic frames are initially made available based on the evidence that all meanings of ambiguous words are initially made available. However, the Concurrent model predicts that only the thematic frame that is most consistent with the context will be pursued.

The Concurrent model predicts that decision times in the ambiguous verb conditions should be faster than those for the unambiguous verb conditions, as in Experiment 1A. This is the syntactic integration effect. In addition, decision

times should reflect semantic integration: the three argument ambiguous condition should be faster than the two argument ambiguous condition because the target is consistent with the context only in the three argument condition. Likewise, decision times for the three argument unambiguous condition should be faster than

those for the two argument unambiguous condition because only the former provides a thematic role for the target. These predictions are summarized in (6a).

In contrast, the most straightforward interpretations of the Interactive and Garden Path models predict the same patterns of effects that the models predicted with the naming task, though for different reasons. Interactive models maintain that syntactic and semantic processors work together to construct a single representation, which both naming and lexical decision would presumably tap. The Garden Path model, in contrast, predicts that the simplest syntactic representation is constructed first (ignoring subcategory information), and it is this initial representation that is presumably being tapped. Thus, there should be no difference between the four conditions (as shown in (6c)) because subcategory information is not yet available. Alternatively, if the task taps a later stage of processing, and subcategory information is available in time to influence the response, one of the two syntactic congruity effects illustrated in (7) should be obtained. If the decision task taps a very late stage in processing, and semantic analysis has also occurred, then the pattern predicted by the serial Interactive model (6b) should be obtained.

6a) Predictions of the Concurrent model:  $3A < 3U = 2A < 2U$

b) Predictions of serial Interactive models:  $3A < 3U = 2A = 2U$

c) Predictions of the Garden Path model:  $3A = 3U = 2A = 2U$

7. Alternative Predictions of the Garden Path model:

a)  $3A < 3U = 2A = 2U$

b)  $3A = 2A > 3U = 2U$

#### *Method.*

**Subjects.** Forty undergraduates at the University of Rochester completed the experiment in partial fulfillment of course requirements, or for a nominal fee. All were native speakers of English.

**Materials.** The auditory contexts used in Experiment 1A were used again here. The target list was modified by generating 24 pronounceable, non-names for the distractor trials. (Non-names were non-words and were not homophonous with any common name or word.) Overall, approximately 30% of the trials were non-name trials. Decision norms were collected on the critical targets without any contexts. The targets used with ambiguous verbs averaged 521 milliseconds and the targets used for unambiguous verbs averaged 518 milliseconds ( $N=8$ ). There were no differences between the two groups.

**Procedure.** The procedure used in Experiment 1A was used again here, except that subjects were told that a string of letters would appear on the screen and they should decide whether or not it was a real name as quickly as possible. Decisions and their latencies were recorded on a button box labeled with "yes" and

"no." If the subject did not respond within three seconds, the program registered a "timeout" and the experiment continued.

### Results.

For each of the critical trials, the lexical decision latency was recorded, along with which button was pressed. Subjects made errors on less than 2% of the critical trials. There were no "time-outs" on the critical trials. Within each condition, correct ("yes") mean response times were computed by subject and by item. Mean response times are displayed in Figure 2, below. Outliers were replaced at 2.5 standard deviations as in Experiment 1A. Approximately 3% of the data were replaced in this way.

Subject and item means were subjected to a 2(list) x 2(verb type) x 2(argument number) ANOVA. The data pattern is strikingly different from that obtained using the pronunciation task. The verb type effect, reflecting syntactic congruity, is still observed, but only in the subject analysis: the ambiguous verb conditions are significantly faster than the unambiguous verb conditions by subjects [ $F(1,38)=28.37, p < .01$ ], but not by items [ $F(1,16)=2.25, p > .10$ ]. The planned comparisons of the two argument ambiguous and unambiguous conditions [ $F(1,38)=10.82, p < .01$ ;  $F(1,16)=2.07, p > .10$ ] and three argument ambiguous and unambiguous conditions [ $F(1,38)=4.85, p < .05$ ;  $F(1,16)=1.49, p > .10$ ] were also significant by subjects but marginal by items.

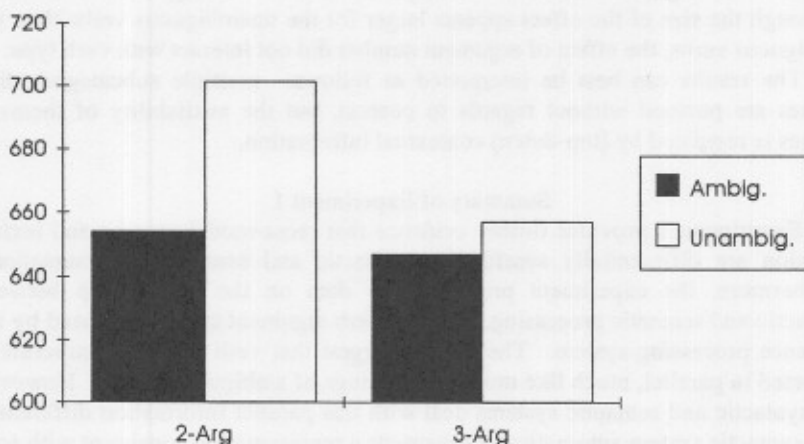


Figure 2. The mean decision latencies for each condition of Experiment 1B are given in milliseconds.

What is striking is the main effect of argument number, with two argument conditions slower than three argument conditions [ $F(1,38)=7.44, F(1,16)=4.69, p < .05$ ]. This reflects semantic integration because there was a thematic role for the target in the three argument condition, but not the two argument condition. Although the effect of argument number did not interact with verb type [ $F_s < 1.0$ ],

it appears numerically larger for the unambiguous verbs. In a planned comparison of the unambiguous verbs, this difference was reliable by subjects [ $F(1,38)=5.21$ ,  $p < .05$ ] and marginally reliable by items [ $F(1,16)=4.19$ ,  $p < .10$ ]. However, the difference was not reliable for the ambiguous verbs [ $F_s < .10$ ].

### *Discussion.*

The pattern of results confirmed that the decision task is sensitive to both syntactic and semantic representations, and that thematic and contextual information are used very early to guide the semantic representation. Evidence that the task is sensitive to thematic information is provided by the unambiguous control conditions. The unambiguous three argument condition was faster than the unambiguous two argument condition because the three argument condition offered a thematic role for the target. The main effect of verb type, seen previously in Experiment 1A, is again evidence of syntactic integration. Responses to the unambiguous verb conditions were comparatively slow because they are not syntactically congruent. By contrast, the ambiguous three argument condition is syntactically congruent on all accounts, and the ambiguous two argument condition is syntactically congruent if the alternative subcategorization frame is available.

The main effect of argument number indicates that only the contextually appropriate thematic frame was pursued. Thus, it was easier to integrate the target in the three argument conditions compared to the two argument conditions. Although the size of the effect appears larger for the unambiguous verbs than the ambiguous verbs, the effect of argument number did not interact with verb type.

The results can best be interpreted as follows: multiple subcategorization frames are pursued without regards to context, but the availability of thematic frames is regulated by (top-down) contextual information.

### Summary of Experiment 1

Experiment 1 provides further evidence that cross-modal naming and lexical decision are differentially sensitive to syntactic and semantic representations. Furthermore, the experiment provides new data on the relationship between syntactic and semantic processing, and how verb argument structure is used by the sentence processing system. The results suggest that verb argument structure is accessed in parallel, much like multiple meanings of ambiguous words. However, the syntactic and semantic systems deal with this parallel information differently. The syntactic system automatically constructs a representation consistent with each subcategorization frame, but the semantic system uses context to select the most likely thematic frame to pursue. In such a system, garden paths would occur only when the thematic system pursued the incorrect interpretation (because the context was misleading or uninformative). In this case, the alternative syntactic frames might be used to identify an alternative analysis.

An alternative line of explanation for the data in Experiment 1 must also be considered. Suppose that the two argument subcategorization and thematic frames of the ambiguous verbs were ruled out, brute force fashion, by the target word. The two argument ambiguous condition would be syntactically congruent, but



implausible, and the three argument ambiguous condition would be syntactically congruent and plausible. Because both would be syntactically congruent there would be no difference in naming times, but the plausibility difference would be reflected in the lexical decision times. This explanation is difficult to rule out, but it is unlikely because there is no reason to think that subjects were forcibly trying to integrate the target word with the contexts. It was assumed that the target would be integrated automatically only if it was congruent. Fully half of the experimental contexts were complete sentences without integrating any probe word so it is unclear why subjects would adopt a strategy of forcibly integrating the probe word. Further evidence against this explanation is provided by Experiment 2. Experiment 2 is a naming experiment that uses the same materials, but probes at time points after the offset of the verb. At late time points, the ambiguous two argument condition is no longer as fast as the ambiguous three argument condition. Thus, it is clear that at 150 and 300 milliseconds post offset, bottom up evidence of the three argument structure does not force that analysis. It is unlikely then, that such a process occur at earlier probe times.

### Experiment 2

Experiment 1 suggests that all subcategorized structures are automatically constructed in parallel, but a single thematic frame is selected using contextual information. Experiment 2 was designed to explore the relationship between syntactic and semantic representations over time by testing the availability of the alternative syntactic frame at different time points. This was accomplished using the naming task in a cross-modal, multi-ISI (inter-stimulus-interval) design. The temporal relationship between the offset of the auditory context and the appearance of the next target was varied from 150 milliseconds prior to offset to 300 milliseconds post offset.

I have assumed that the naming data and the lexical decision data reflect the same time point in processing. Thus, the results of Experiment 1 demonstrate that a single interpretation has been selected **before** the syntactic ambiguity is resolved. However, it is also possible that the naming task captures the sentence processor at an earlier point than does lexical decision -- that, in fact, the semantic processor cannot develop an interpretation until a single syntactic structure is passed up by the syntactic processor. This possibility calls into question the task difference, itself, because it follows that if the naming response was slowed down the task would be sensitive to semantic integration. The latter possibility is ruled out by the naming data from the current experiment. We will see that even 450 milliseconds later (300 milliseconds post-offset), there is no evidence of a semantic effect with the naming task.

This leaves the question of how the subcategorization ambiguity is resolved. According to the Concurrent model, once syntactic and semantic representations are developed, they are compared, and inconsistent representations are discarded. This leads to the prediction that argument number and ISI should interact when the verb is ambiguous. Specifically, the ambiguous two argument condition should be fast at short ISI's and slow at long ISI's with the naming task. (It should be slow at all ISI's using lexical decision.) On the other hand, if syntactic parallelism is

maintained until there is bottom-up evidence for one subcategorization over the other, one would expect the ambiguous two argument condition to remain fast in the naming task, because integration of the probe word should constitute bottom-up evidence for the three argument subcategorization.

The data pattern for the ambiguous verbs that is predicted by the Concurrent model is identical to the pattern predicted if naming is sensitive to semantic congruity at long ISI's. Therefore, the control conditions are again crucial. If naming is truly sensitive only to syntactic congruity, argument number should not interact with ISI when the verb is unambiguous. That is, the unambiguous three argument condition should not be faster than the unambiguous two argument condition at long ISI's. However, if the source of the task difference between naming and lexical decision is in the relative timing of the response -- and naming was tapping an earlier representation than lexical decision in Experiments 1 and 2 -- then the pattern of naming responses should mimic the lexical decision task at long ISI's, and the unambiguous three argument condition should be faster than the unambiguous two argument condition.

#### *Method.*

**Subjects.** Eighty undergraduate students from the Ohio State University served as subjects, 20 in each of the ISI conditions. All were naive to the experimental hypothesis and were native speakers of English. Subjects participated to fulfill part of their course requirement in introductory psychology.

**Materials.** The materials from Experiment 1 are used again here, although they were spoken by a different person, re-randomized and assigned different targets.

**Procedure.** The auditory sentence contexts were digitized and edited as in Experiment 1. As before, sync tones were set at approximately 150 milliseconds before the offset of the last word in the auditory context. However, the temporal relationship between the sync tone and the presentation of the target word was manipulated between subjects so that there were 4 ISI conditions. The target was presented at the sync tone for the "-150 ISI" group, 150 milliseconds after the sync tone for the "0 ISI" group, 300 milliseconds after the sync tone for the "150 ISI" group, and 450 milliseconds after the sync tone for the "300 ISI" group.

The only other procedural change was that the fastest and most accurate subject in each task was awarded a \$10 prize. This incentive produced somewhat faster response times than were seen in Experiment 1.

#### *Results.*

The results are summarized in Figure 3. Outliers were replaced at 2.5 standard deviations and subject and item means were computed as in the previous experiments. These means were first subjected to a 2(list/item group) x 2(verb type) x 2(argument number) x 4(ISI) analysis of variance. In the subject analysis, list and ISI were between factors and verb type and argument number were within factors. In the item analysis, item group and verb type were between factors, while argument number and ISI were within.

The predicted effect of verb type (with the ambiguous verb conditions faster than the unambiguous verb conditions) was obtained in the subject analysis [ $F(1,72)=13.04, p < .001$ ], but was not reliable in the item analysis [ $F(1,16)=1.93, p > .10$ ]. Importantly, this effect did not interact with ISI [ $F(1, 72) < 1.0$ ], demonstrating that the syntactic congruity effect was maintained across time. In addition, there was no main effect of argument number [ $F(1, 72) < .10$ ], which would have reflected a semantic congruity effect. The main effect of ISI was reliable in the item analysis, but not in the subject analysis [ $F(3,72)=1.39, p > .10$ ;  $F(3,14)=16.14, p < .01$ ].

Unfortunately, the predicted three-way interaction between verb type, argument number, and ISI was not obtained [ $F(1, 72) < 1.0$ ]. This interaction was predicted because argument number and ISI should interact for ambiguous verbs, but not for unambiguous verbs. Although the predicted pattern was obtained at the last three probe positions, with the ambiguous two argument condition gradually becoming less available, the ambiguous two argument condition was inexplicably (and non-reliably) slower than the ambiguous three argument condition at the first probe position. Thus, instead of the predicted three-way interaction, only a two-way interaction between verb type and argument number was obtained [ $F(1, 72)=7.08, p < .01$ ;  $F(1,16) = 6.32, p < .05$ ].

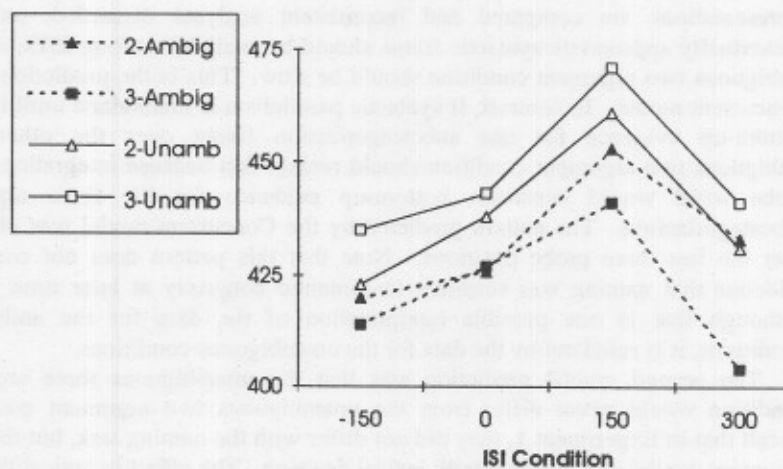


Figure 3. The mean naming latencies for each condition of Experiment 2 are given in milliseconds.

Because the model makes predictions about differences between ISI conditions, separate 2(list) x 2(verb type) x 2(argument number) ANOVAs were done on each ISI group. At the earliest ISI (-150), the data was quite noisy, and no main effects or interactions reached significance. By the offset of the verb (0 ISI), the main effect of verb type was reliable by subjects [ $F(1,18)=5.44, p = .01$ ], but not by items [ $F(1,16)=1.48, p > .10$ ]. Shortly after verb offset, at 150

ISI, the effect of verb type was only marginally reliable by subjects [ $F(1,18) = 3.99, p < .10; F(1,16)=1.53, p > .10$ ], and it failed to reach significance completely by 300 milliseconds post offset [ $F(1,18)=2.48, F(1,16)=2.26, p > .10$ ]. At the last ISI, there was also an interaction between verb type and argument

number [ $F(1,18)=7.18, F(1,16)=5.83, p < .05$ ]. There was never an effect of argument number [ $F_s < 1.0$ ].

Planned comparisons of the ambiguous and unambiguous three argument conditions demonstrated that the ambiguous three argument condition was reliably faster by subjects and marginally faster by items at 150 milliseconds post-offset ( $\alpha = .05$ ). The difference was reliable in both subject and item analyses at 300 milliseconds post offset. The unambiguous three argument condition never differed reliably from the unambiguous two argument condition in either the subject or the item analyses. The only ISI at which the ambiguous three argument condition was reliably faster than the ambiguous two argument condition was the 300 ISI condition.

#### *Discussion.*

Experiment 2 used the naming task to examine the state of the syntactic representation(s) at various points in time. There were two crucial predictions, one theoretical and one methodological. First, if the syntactic and semantic representations are compared and inconsistent analyses discarded, only the contextually appropriate syntactic frame should be available at long ISI's, and the ambiguous two argument condition should be slow. This is the prediction of the Concurrent model. In contrast, if syntactic parallelism is maintained until there is bottom-up evidence for one subcategorization frame over the others, the ambiguous two argument condition should remain fast because integration of the probe word would constitute bottom-up evidence for the three argument subcategorization. The pattern predicted by the Concurrent model was obtained over the last three probe positions. Note that this pattern does not constitute evidence that naming was sensitive to semantic congruity at later time points. Although that is one possible interpretation of the data for the ambiguous conditions, it is ruled out by the data for the unambiguous conditions.

The second crucial prediction was that the unambiguous three argument condition would never differ from the unambiguous two argument condition. Recall that in Experiment 1, they did not differ with the naming task, but the three argument condition was faster with lexical decision. The effect in lexical decision was attributed to the task's sensitivity to semantic congruity, because the unambiguous three argument verbs allowed a third thematic role that was consistent with the semantic features of the target. However, there is an alternative explanation consistent with the Garden Path model that must be ruled out. The alternative is that lexical decision task taps processing at a later stage than does naming. Under this account, both naming and lexical decision are sensitive to semantic congruity in principle, and semantic effects would be seen in naming if there were enough time for the semantic information to become available. To rule out this explanation, the unambiguous three argument condition must remain slow at all ISI's. As seen in Figure 3, this pattern was obtained; there

was no difference between the unambiguous three argument condition and the unambiguous two argument condition, even at the longest ISI. This demonstrates that naming was insensitive to thematic congruity in this paradigm.

Somewhat surprisingly, no clear pattern emerged at the first ISI, and this appeared to disrupt the predicted three-way interaction. Although the -150 ISI was used in Experiment 1, the data pattern found in Experiment 1 is seen here at verb offset. This discrepancy is probably due to the relatively short naming latencies observed here compared to those in Experiment 1. Thus, responses at verb offset are presumed to reflect the same stage of processing observed in Experiment 1. As before, both subcategorization frames for the ambiguous verbs were available. Unlike Experiment 1, there does appear to be a difference between the unambiguous control conditions -- the three argument condition is slower, not only here, but at each ISI. This difference between the control conditions was not predicted by any of the models, and was not reliable at any ISI.

The most interesting data is from the longer ISI conditions. By 300 milliseconds post-offset, only the contextually appropriate subcategorization frame was available, and the predicted interaction between verb type and argument number was obtained. Crucially, there is never an effect of argument number, which would have reflected semantic congruity. The apparently gradual decrease in the availability of the alternative subcategorization frame is consistent with a gradual decay in its activation level once the appropriate syntactic representation has been successfully matched to the semantic representation.

### General Discussion

This set of experiments used cross-modal naming and lexical decision in the Integration Paradigm to explore how subcategory and thematic information is used by the sentence processing system. Because this combinatory lexical information must be used by the sentence processing system at some point, investigations into how they are used will also provide insight into the relationship between syntactic and semantic processing. A Concurrent model of sentence processing was proposed, and some of its predictions were tested against the competing interactive and modular models. The evidence provided here supported the Concurrent model.

The Concurrent model of sentence processing maintains that, when a verb is recognized, all of its subcategory and thematic information is accessed. The syntactic processing system uses the subcategory information to construct all the subcategorized structures in parallel, without consulting semantic or contextual sources of guidance. Meanwhile, the semantic processing system uses contextual information, along with any preliminary output from the syntactic system, to select the most likely interpretation. Once semantic and (parallel) syntactic representations are constructed, they are compared to eliminate inconsistent analyses. This is similar to other processing models in which multiple syntactic structures are proposed in parallel (e.g., Crain & Steedman, 1985; Gorrell, 1989 & 1991; Hickok, 1993; MacDonald, 1993; McElree, 1993), except that the Concurrent model makes the explicit claim that a single interpretation is sometimes constructed before the syntactic system has identified the appropriate

structure. This claim was supported by the lexical decision results of Experiment 1B, especially when compared against the naming results of Experiment 2. Subcategorization information, thematic information, and contextual information were all used very early, but only subcategorization information influenced the

initial syntactic representations. All three sources of information were used to guide the semantic interpretation. This pattern of data is consistent with Parallel Syntax models that allow a single semantic representation to be constructed before a single syntactic analysis is selected.

The evidence from both experiments suggests that, under appropriate experimental conditions, naming is sensitive to syntactic representations and relatively insensitive to semantic representations. In contrast, lexical decision is sensitive to both syntactic and semantic representations. However, this task difference is not as reliable as one would hope. For instance, Gorrell (1991) found that lexical decision was *insensitive* to animacy violations, and Duffy and colleagues (e.g., Duffy et al., 1989) have found that naming was *sensitive* to plausibility. It is important to note, particularly when comparing these experiments to similar studies, that the targets were presented to the subjects without warning, and prior to the offset of the final context word. Many studies have presented a warning signal before the target or offset the target from the context by half a second or more, altering the nature of the task. Furthermore, the paradigm may be limited in utility to cross-modal presentation. Experiments in my own laboratory that have used visual presentation have produced noisy results without clear task differences.

Nonetheless, the Integration Paradigm may prove useful because it has the important capability of mapping changes in representations over time. This feature was exploited in Experiment 2, to examine changes in the syntactic representations. The results suggested that the contextually inappropriate subcategorization frame gradually became less available. I have suggested that the alternate syntactic representation could be used to recover from a garden path, but as yet I have offered no evidence of this. Further research is necessary to determine if, in fact, alternative subcategorization frames do serve as a mechanism for recovery from a garden path by manipulating the point at which disambiguating information becomes available. I predict that recovery would be more efficient if disambiguating material appeared while the corresponding syntactic representation was still available.

#### References

- Bates, E. & MacWhinney, B. (1982) Functionalist approaches to language acquisition. In *Language acquisition: The state of the art* (E. Wanner & L.R. Gleitman, editors) Cambridge: Cambridge University Press.
- Boland, J.E. (1991) *The use of lexical information in sentence processing*. Doctoral dissertation. The University of Rochester, Rochester, NY.
- Boland, J.E. (1993) The role of verb argument structure in sentence processing: Distinguishing between syntactic and semantic effects. *Journal of Psycholinguistic Research*, 22, 109 - 132.

- Clifton, C., Jr., Frazier, L. & Connine, C. (1984) Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, **23**, 696-708.
- Crain, S. & Steedman, M. (1985) On not being led up the garden path: The use of context by the psychological syntax processor. In *Natural language parsing: Psychological, computational, and theoretical perspectives* (D.R. Dowty, L. Karttunen, & A.M. Zwicky, editors) Cambridge: Cambridge University Press.
- Duffy, S. A., Henderson, J. M., & Morris, R.K. (1989) Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 795-801.
- Ferreira, F. & Henderson, J. M. (1990) The use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**, 555-568.
- Forster, K.I. (1979) Levels of processing and the structure of the language processor. In *Sentence processing: Studies presented to Merrill Garrett* (W.E. Cooper & E.C.T. Walker, editors) Hillsdale, NJ: Erlbaum.
- Frazier, L. (1987) Theories of syntactic processing. In *Modularity in Knowledge Representation and Natural Language Processing* (J.L. Garfield, editor) Cambridge, MA: MIT Press.
- Frazier, L. (1989) Against lexical generation of syntax. In *Lexical Representation and Process* (W.D. Marslen-Wilson, editor) Cambridge, MA: MIT Press.
- Frazier, L. & Clifton, C., Jr. (1989) Successive cyclicity in the grammar and the parser. *Language and Cognitive Processes*, **4**, 93-126.
- Garnsey, S.M., Tanenhaus, M.K., & Chapman, R.M. (1989) Evoked potentials and the study of sentence comprehension. *Journal of Psycholinguistic Research*, **18**, 51-60.
- Gorrell, P. (1989) Establishing the loci of serial and parallel effects in syntactic processing. *Journal of Psycholinguistic Research*, **18**, 61-74.
- Gorrell, P. (1991) Informational encapsulation and syntactic processing. *NELS proceedings*, **21**.
- Hagoort, P. Brown, C., & Groothusen, J. (1993) The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, **8**, 439-484.
- Hickok, G. (1993) Parallel parsing: Evidence from reactivation in garden-path sentences. *Journal of Psycholinguistic Research*, **22**, 239 -250.
- Kurtzman, H.S. (1989) Locating Wh-traces. In *The MIT Parsing Volume, 1988-89* (C. Tenney, editor).
- Marslen-Wilson, W.D. (1987) Functional parallelism in spoken word-recognition. *Cognition*, **25**, 71-102.
- Marslen-Wilson, W.D. & Tyler, L.K. (1987) Against Modularity. In *Modularity in Knowledge Representation and Natural Language Processing* (J.L. Garfield, editor), Cambridge, MA: MIT Press.

- MacDonald, M.C. (1993) The interaction of lexical and syntactic ambiguity. *Journal of Memory & Language*, **32**, 692 - 715.
- McElree, B. (1993) The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language*, **32**, 536-571.
- Mitchell, D.C. (1989) Verb-guidance and other lexical effects in parsing. *Language and Cognitive Processes*, **4**, 123-154.
- Neville, H., Nicol, J., Barss, A., Forster, K.I., & Garrett, M.F. (1991) Syntactically based sentence processing classes: Evidence from even-related brain potentials. *Journal of Cognitive Neuroscience*, **3**, 151-165.
- Osterhout, L. & Holcomb, P.J. (1992) Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory & Language*, **31**, 785-806.
- Rayner, K., & Morris, R.K. (1991) Comprehension processes in reading ambiguous sentences: Reflections from eye movements. In *Understanding word and sentence* (G.B. Simpson, editor), North-Holland: Amsterdam.
- Tabossi, P. (1988) Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language*, **27**, 324-340.
- Tanenhaus, M.K., Carlson, G.N., & Trueswell, J.C. (1989) The role of thematic structures in interpretation and parsing. *Language and Cognitive Processes*, **4**, 211-234.
- Van Petten, C. & Kutas, M. (1987) Ambiguous words in context: An event-related potential analysis of the time course of meaning activation. *Journal of Memory and Language*, **26**, 188-208.
- Waltz, D.L. & Pollack, J.B. (1985) Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, **9**, 51-74.



## Appendix

The critical materials used in the two experiments are listed here. The first version of each item is the two argument-condition and the second is the three argument-condition.

### A. Ambiguous Verbs.

1. Which salad/baseball did Jenny toss
2. Which chapter/letter did Howard write
3. Which prison sentence/fancy dessert did Henry serve
4. Which dark alley/salt shaker did Linda pass
5. What dress/fee did Mrs. Smith charge
6. What town/package did Mr. Simpson leave
7. Which victims/seat did Martin save
8. What kind of tantrum/frisbee did Becky throw
9. Which excuse/gift did Robyn buy
10. Which new magazine/bedtime story did Alice read

### B. Unambiguous Verbs.

1. Which necklace did Nancy inspect/describe
2. Which poem did Martha finish/dedicate
3. Which friend did Leonard insult/introduce
4. Which quote did Kathy underline/explain
5. Which pie did Mrs. Jones smell/recommend
6. Which package did Cindy open/deliver
7. Which notebook did Patty damage/return
8. Which secret recipe did Nancy follow/entrust
9. Which hotel did Mr. Peterson examine/mention
10. Which task did Larry despise/demonstrate

# The Influence of Orthography and Sentence Constraint on the Processing of Nouns in Japanese

Kim Darnell<sup>1</sup>, Julie Boland<sup>2</sup>, Mineharu Nakayama<sup>3</sup>

**Abstract:** Utilizing a word-by-word reading paradigm, we investigated the role of orthographic familiarity in the processing of Japanese nouns by comparing the reading times of words that were *kanji* dominant (the *kanji* form is preferred by native speakers), *kana* dominant (the *kana* form is preferred), and orthographically neutral (both forms are equally acceptable). Target words appeared in *kana* or *kanji*, and were embedded in highly constraining (Experiment 1) or unconstraining (Experiment 2) carrier sentences. The results suggest that orthography does not affect reading time unless the sentence is highly constraining, in which case the most familiar orthography is faster.

For the most part, research on visual word recognition and sentence processing has focused on English and other alphabetic languages. Much less is known about how Japanese is processed. However, Japanese is an interesting language to investigate because, unlike English, it is head-final, allows pro-drop, and has three distinct orthographic systems: the logographic *kanji* and two *kana* syllabaries, *hiragana* and *katakana*. *Kanji* are logographic characters used to indicate meaning for content words, such as nouns and roots of verbs, adjectives and some adverbs. *Hiragana* is used for function words, the inflectional endings of verbs, adjectives and adverbs, and some nouns. *Katakana* is used primarily for representing loan words and onomatopoeic expressions. Thus, a single sentence may be composed of a mixture of all three orthographic systems.

The current paper investigates recognition of *kanji* and *hiragana* by varying the orthography and orthographic familiarity of target words in sentence contexts. In doing so, we hoped to discover how these variables affect reading time in typical Japanese sentences.

*Kanji* and *kana* differ in some important ways. The phonological readings for *kanji* are dependent on several factors, including the origin of the word in which a character appears, if the character is part of a compound, and the sentential context (see Figure 1). For *kana*, conversely, the readings are

\*This research was made possible through a fellowship funded by the Center for Cognitive Science, the Department of Linguistics, and the Department of East Asian Languages and Literatures, all of the Ohio State University. The authors also wish to thank Keith Johnson, Mary Beckman, and Rob Fox for their guidance and advice, and Teruaki Hirano, Hiroko Butler, and Tomokazu Umeki for their native speaker judgements. All questions and comments concerning this paper should be addressed to the first author, c/o the Department of Linguistics, 1712 Neil Ave., Columbus, Ohio, 43210, or made via electronic mail to darnellk@ling.ohio-state.edu.

<sup>1</sup>Department of Linguistics, the Ohio State University

<sup>2</sup>Department of Linguistics and Department of Psychology, the Ohio State University

<sup>3</sup>Department of East Asian Languages and Literatures, the Ohio State University

completely invariant, each character representing a single, distinct mora<sup>4</sup> (Morton & Sasanuma, 1984). Furthermore, *kanji* are associated with particular meanings, while *kana* possess no inherent semanticity (Aoki, 1990; Elman et al., 1981; Hatta, 1978).

Character	<u>in word of Chinese origin</u>	<u>in word of Japanese origin</u>
性	sei, shou	saga
<u>in different character compounds</u>		
性犯罪	seihanzai	'sex crime'
性分	shoubun	'disposition'
<u>same character, different context</u>		
性	sei	'sex'
性	saga	'one's custom'

**Figure 1.** The different phonological readings of *kanji*

The differences between *kanji* and *kana* might have consequences on how the two types of orthographies access the lexicon. Of particular interest here are differences that might impact the speed of word recognition, and thus, reading time. For example, many have suggested that *kanji* access the lexicon by their physical form alone, while *kana* require the reader to recode phonologically before access is possible (Allport, 1979; Goryo, 1987; Inoue et al., 1979; Kimura, 1984; McCusker et al., 1981; Morton & Sasanuma, 1984). In order to test this claim empirically, one might assume that access via a direct, visual route is faster than access via the indirect, phonological route. This assumption predicts that words written *kanji* will be accessed more quickly than words written in *kana*. However, Besner & Hildebrandt (1987) found that words normally written in *katakana* were named more quickly when presented in *katakana* than *katakana* nonwords and words that normally appeared in *kanji* presented in *katakana*. They concluded that common *kana* can access the lexicon directly. Similar conclusions were drawn from empirical work by Hirose (1984, 1985) and Sasanuma and colleagues (Sasanuma et al., 1988) utilizing familiar *hiragana* words.

Assuming that direct access is faster than indirect access warrants caution on other grounds. When comparing across different visual stimuli, it is difficult to control early visual analysis time. Logographic characters can be very complex,

<sup>4</sup> A mora is defined as a sound unit which is produced for a certain length of time, and is sometimes equivalent to a syllable. In Japanese, it may be composed of a vowel, a consonant and a vowel, a single nasal, or a pinate consonant.

and it might take longer to perceive the relevant features of a word written in *kanji* as compared to one presented in *kana*. In this event, any difference in recognition time predicted by the different routes of access could be eliminated.

One goal of the current project is to explore to what degree *kanji* and *kana*

may be interchanged in text without significantly affecting processing time. A better understanding of the time-course of word recognition in the two orthographies will facilitate studies of Japanese sentence processing, particularly those using lexical ambiguity. To this point, such work has been inhibited due to the fact that the writing system has a specific means of distinguishing like tokens: words that would be ambiguous if written only in *kana* are instead written in *kanji* or a combination of *kanji* and *kana* (Aoki, 1990; Sasanuma et al., 1977).

But recall the evidence that familiar *katakana* and *hiragana* words were recognized faster than words that were unfamiliar in *kana* form. Orthographic familiarity, rather than orthography, might be the best determinate of reading speed. Interviews we conducted with native speakers show that within the lexical category of nouns, there are in fact words for which the *hiragana* form is dominant (i.e. native Japanese feel that the word most commonly appears in *hiragana* and is the most acceptable in this form), as well as those which are orthographically neutral. If we posit that the familiarity of the visual form, not whether it is logographic or syllabic, is the key to speed of lexical access, we can make some interesting predictions. The processing of *kana* dominant nouns should be slowed if they are presented in *kanji*. Moreover, nouns with no orthographic bias should display little difference in speed of facilitation between the *kanji* and *kana* forms. Such results would be particularly informative since other factors which affect processing speed, such as frequency (how often one is exposed to the word regardless of orthography) and concreteness (how salient an image one can create in association with the word), are identical for the *kanji* and *kana* forms of any given word, and therefore controlled variables.

The current study investigates the role of orthographic familiarity on the speed of lexical access by comparing response times for the *kanji* and *hiragana* forms of words with different script dominances embedded in context-biased sentences, where each target is preceded by lexical associates (Experiment 1), and non-biased sentences, where each target is semantically congruent but unpredictable (Experiment 2). The manipulation of context is likely to have two effects. The context-biased condition was designed to minimize any ambiguity for *kana* targets by insuring that, in the event that the target word has homophones, there is sufficient degree of contextual priming to eliminate the *kana* form being interpreted as having a meaning other than the one given by the corresponding *kanji*. In doing so, however, we make the target words highly predictable compared to the non-biased condition. Predictability may well interact with familiarity if a biasing context leads the reader to expect the dominant orthographic form of a word.

## Experiment 1

### METHOD

**Subjects** Twenty native speakers of Japanese currently living in the greater Columbus area were used as subjects. All participants were between the ages of 18-40, were educated in Japan through high school, had lived in the U.S. for no more than five years, and had normal or corrected-to-normal vision. Each subject was paid a nominal fee for participating.

**Apparatus** The program for this experiment was written on the Macintosh KANJI TALK operating system and was presented on a Macintosh SEII monochrome screen. A customized response box was used to collect subject responses.

**Stimuli** The stimuli for this project came from three distinct groups of words: *kanji* dominant, where the *kanji* form is considered by native speakers to be the most familiar; *kana* dominant, where the *hiragana* form is most familiar; and orthographically neutral, where both forms are equally familiar. Data on words and orthographies were solicited via a questionnaire from fifteen native speakers representative of the intended subject pool. None of these individuals acted as subjects in this study.

All non-nominal items were eliminated from the collection of potential stimuli. The remaining words were then collapsed into a single list. This list was then redistributed for ranking of frequency of appearance of *kanji* and *kana* forms in everyday written material (1 = never 2 = very rarely 3 = rarely 4 = sometimes 5 = often 6 = very often 7 = all the time). Pairs in which the *kanji* form had a familiarity average of at least 2.5 points higher on the seven point scale than the *hiragana* form were considered *kanji* dominant, with the opposite requirement for *kana* dominant words. Orthographically neutral pairs were those in which the average scores for both scripts were within .5 of each other. All words were then rated for concreteness to control potential lateral differences in processing<sup>5</sup>; the characters with the highest ratings were given preference in their dominance category. The 10 pairs which best met both the frequency and concreteness criteria were chosen for each stimuli group. Group means are listed by dominance condition and orthography in Table 1 below.






	<i>Kanji</i> Dominant	<i>Kana</i> Dominant	Ortho Neutral
<i>Kanji</i>	6.88	2.64	5.33
<i>Kana</i>	2.61	5.83	5.41

**Table 1.** Mean familiarity ratings for *kanji* and *kana* forms of stimuli by dominance

Each stimulus was embedded in a sentence that contained "lexical associates"-other words which are strongly associated with the given item. For instance, consider (1). The target word is *rousoku* 'candle', its lexical associates are *tanjoubi* 'birthday', *keeki* 'cake', and *tatsu* 'to stand'. All of the associates appear before the target, affording a degree of contextual priming. In the event that the target word has homophones, this should eliminate the possibility of the stimuli written in *kana* being interpreted as having a meaning other than the one given by the corresponding *kanji*.

<sup>5</sup> Some tachistoscopic studies have suggested lateral preferences for *kanji* and *kana*, (Hatta, 1976, 1977, 1978; Hirata and Osaka, 1967; Sasanuma et al., 1977). However, these claims are contrary to English based experiments concerning the processing of abstract lexical items like adjectives and verbs (Elman et al., 1981) and concrete words like nouns (Caplan et al., 1974; Day, 1977; Ellis and Shepard, 1974; Hines, 1976, 1977; Shanon, 1979). Furthermore, Ohnishi and Hatta (1980) argue that the degree of concreteness of the *kanji* itself may control which hemisphere is dominant in processing. To avoid potential complications related to this issue, we elected to use only nouns that refer to concrete, easily visualized items.

- (1) target: *rousoku* 'candle'  
 lexical associates: *tanjoubi* 'birthday', *keeki* 'cake', *tatsu* 'to stand'

<i>tanjoubi-no</i>	<i>keeki-no</i>	<i>ueni</i>	<i>taterareta</i>	<i>rousoku-wa</i>
birthday-Gen	cake-Gen	top on	stand-pass-past	candles-Top
				

<i>kireini</i>	<i>maru-o</i>	<i>egaitaita</i>
pretty	circle-Acc	arrange-pass-past

'The candles placed on the top of the birthday cake were arranged in a pretty circle.'

A questionnaire like that distributed by Tabossi (1988) was used to solicit associate words for each stimulus from 20 native Japanese speakers. The two to three most frequently suggested associates that could be used to produce a semantically congruous sentence were selected for each target. To minimize variables related to syntactic processing, every attempt was made to place the targets in the same syntactic position in each sentence, namely the direct object position. In some cases, however, the most acceptable place for the target was in the subject position; due to the head-final nature of Japanese, it was still possible to place the appropriate associates before the target in these instances.

Aside from the orthographic manipulation of the stimuli, the experimental sentences were presented in characters consistent with convention; the same carrier sentence was used for both the *kanji* and *kana* form of each stimulus.

There were two experimental lists. List 1 contained five stimuli from each dominance category in the *kana* form and five in the *kanji* form, while List 2 contained the same words in the opposite forms. To compensate for any lexical priming which might occur due to different stimuli having similar associates, or sentences containing words or *kanji* characters which might influence the reading speed of critical trials, each list had two orders. This allowed for a post-hoc analysis of order of sentence presentation, so that any such priming effects could be considered in the final interpretation of the data. Thirty-five distractor sentences of various types were added to each list to prevent subjects from developing a strategy of response to critical trials. (A full set of materials is available from the first author.)

### PROCEDURE

Subjects were seated in front of the computer and shown the YES and NO keys on the response box; the YES key was always under the dominant hand. The sentences were presented in a self-paced, modified word-by-word format; subjects proceeded from word to word by pushing the YES key. Each word appeared in the center of the screen in 24 point font, surrounded by a one millimeter rectangle frame.

"Words" consisted of a noun and a particle, a modified noun and particle, an adjective or adjectival noun (possibly modified by an adverb) and inflectional ending, or a verb. The critical stimuli were always presented as a noun and a particle (see Figure 2).

	<u>Japanese</u>	<u>Romanization</u>	<u>Gloss</u>
Noun and particle	お母さんが	<i>okaasan ga</i>	mother-Nom
Modified noun and particle	髪の毛は	<i>kami no ke wa</i>	hair-Top
Adjectival noun with inflection	いろいろな	<i>iroiro na</i>	various
Verb	歩く	<i>aruku</i>	walk

**Figure 2.** Examples of words used in modified word-by-word task

At the end of each sentence, the subject saw a lexical item which he or she had to identify as either being or not being in that trial by pressing the YES or NO key; these probe words differed from the previously described phrases in that they were not followed by particles, and had the word "judge" above the frame. In critical trials, the probe words were chosen from among the lexical associates of the stimulus. This task was used as an accuracy filter, to make sure that subjects were reading each phrase presented to them.

One third of the trials were followed by comprehension questions to encourage subject attentiveness. To familiarize the subjects with the self-paced reading procedure, ten practice trials preceded the experimental trials.

After the experiment, each participant was given two questionnaires. The first tested the subject's ability to read the *kanji* forms of the 30 stimuli, to insure that he or she was actually capable of processing each stimulus in that orthography. The second questionnaire was identical to the one used to gather familiarity data for each stimulus (see *Stimuli*, above). Subjects' frequency ratings for the *kanji* and *kana* forms for each item were averaged and compared with the initial ratings to make sure there were no significant discrepancies.

**Design** This project had a design of 3(dominance) x 2(orthography) x 2(list) x 2(order), with dominance and orthography being within subject factors, and list and order being between subject factors.

## RESULTS

Subjects had to be at least 90% accurate in probe word identification task for their reading times to be included in the experimental data. Five subjects, not included in the count of 20 given above, did not meet this criterion. Subjects who could not read a *kanji* form that had appeared in their version of the experiment had their reading times for that form eliminated from the data, since no lexical access could have taken place. These errors accounted for 22% of the *kanji* trials overall, but all fell within the *kana* dominant condition, which had a resulting error rate of 66%.

Means were taken of the reading times for the *kanji* and *kana* forms of each stimulus across subjects and items for each experimental condition (see Figure 3 for subject means). A range of acceptable reading times for each form of each item was defined as 2.5 standard deviations above and below the item mean.

Reading times for an item which fell outside of this range were replaced with the cut-off value. Three percent of the values were replaced in this way. The resulting sets of reading times were then analyzed by means of a three factor Analysis of Variance (ANOVA). A one-way post-hoc analysis (planned comparison) was also performed on the subject and item means for the *kanji* and

*kana* forms in each dominance condition.

We found that subjects were able to read a word faster when it was presented in its dominant orthography, with the interaction of orthographic dominance and orthography of presentation significant by subjects [ $F_1(2,34) = 4.946, p < .05$ ] and by items [ $F_2(2,34) = 4.034, p < .05$ ]. As anticipated, there was no difference between the reading times in the neutral condition for *kanji* and *kana* by subjects [ $F_1(1,17) = .029, p > .05$ ] or by items [ $F_2(1,17) = .269, p > .05$ ]. There was also, however, no significant difference between orthographies in the *kanji* dominant condition by subjects or items [ $F_1(1,17) = 1.306, p > .10$ ;  $F_2(1,17) = 1.998, p > .05$ ]. The reading times for *kanji* and *kana* were reliably different in *kana* dominant condition by subjects [ $F_1(1,17) = 10.203, p < .05$ ], but not by items [ $F_2(1,17) = .520, p > .05$ ].

Across orthographies, there was a main effect of orthographic dominance by subjects [ $F_1(2,34) = 4.278, p < .05$ ], as well as a marginal effect of orthography of presentation [ $F_1(1,17) = 3.164, p < .10$ ]. By items, however, there was only a marginal effect of dominance [ $F_1(2,34) = 2.796, p < .10$ ]. *Kana* dominant words were read more slowly than *kanji* dominant words, which were read more slowly than orthographically neutral words.

The results of the familiarity ratings are summarized in Table 2 below.

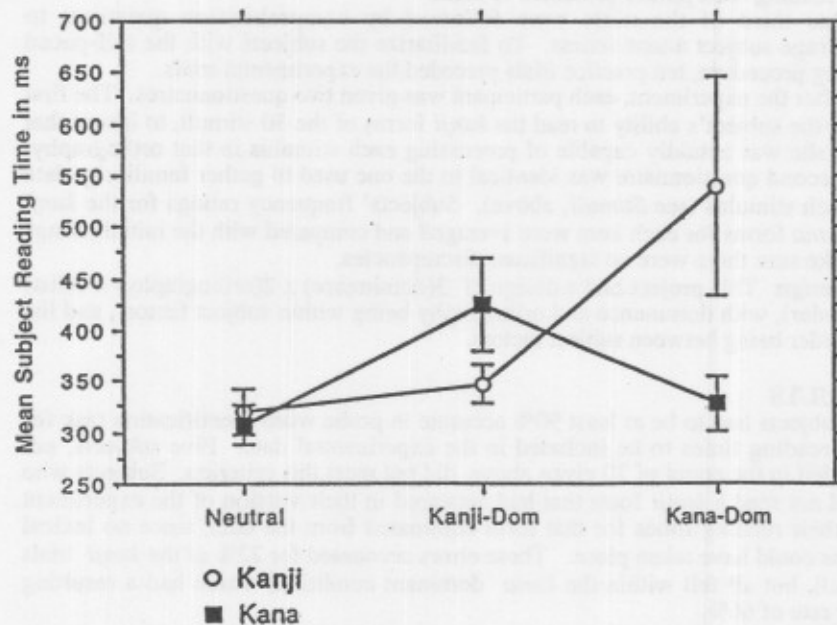


Figure 3. Mean reading time by orthography and orthographic dominance for Experiment 1, with standard error bars



	<i>Kanji</i> Dominant	<i>Kana</i> Dominant	Ortho Neutral
<i>Kanji</i>	6.74	2.40	5.03
<i>Kana</i>	1.85	6.10	4.67

**Table 2.** Experiment 1 mean familiarity ratings for *kanji* and *kana* forms of stimuli by dominance

## DISCUSSION

Our results bring two points to the forefront. First, despite the physical differences between *kanji* and *kana*, it is possible for the two orthographies to be processed at the same rate when familiarity is controlled (as in the orthographically neutral condition). Second, the orthographic form that is most familiar is processed more readily than a less familiar form when the context is highly constraining. This is true even if the less familiar form is logographic and has more morphological content than the familiar script.

Still, there are two limitations to the conclusions we can draw. First, reading times for the *kanji* forms of words in the *kana* dominant condition appear drastically slower than *kana* times in the *kanji* dominant condition and have a huge range of standard error. This is undoubtedly due to the difficulty subjects had reading the unfamiliar *kanji* forms; further evidence for this difficulty is the large number of missing values in the cell caused by subjects' inability to read the *kanji* forms of certain words. In fact, post-hoc tests revealed that the difference between *kanji* and *kana* in the *kana* dominant condition was significant by subjects but not significant by items. Second, there is the possibility that the remarkably similar reading speeds for the *kanji* and *kana* forms of orthographically neutral words might be the result of the strongly biased context in which the stimuli were embedded. Somehow this biasing might neutralize the semantic advantage of the *kanji* form by heavily priming the stimulus and making it predictable, regardless of orthography of presentation. Conversely, the familiarity effect might be caused by the contextual bias. It is possible that the contexts are priming specific orthographies rather than abstract concepts. To investigate this issue, we performed a second experiment in which the stimuli were embedded in semantically plausible, but non-biased contexts.

## Experiment 2

### METHOD

**Subjects** There were twenty participants, different from those in Experiment 1, but from the same subject pool. Each was paid a nominal fee for their involvement.

**Apparatus** The same equipment was used as in Experiment 1.

**Stimuli** The targets from Experiment 1 were used in sentences that were not semantically biased toward the targets, but were restrictive toward the intended meaning (to eliminate potential interference from homophones). Consider example (2), below. As in (1), the stimulus is *rousoku* 'candle'. In this carrier, however, there are no lexical associates or other clues in the sentence which lead the reader to expect the stimulus in question.

(2) target: *rousoku* 'candle'

*Yamamoto-san-wa*      *chiisana*      *kawaii*      *nuigurumi-o,*  
Miss Yamamoto-Top      little      cute      stuffed animals-Acc

*soshite*      *ruumumeito-no*      *Morii-san-wa*      *rousoku-o*  
and      roommate-Gen      Miss Morii-Top      candles-Acc

*atsumeteita*  
collect-past

'Miss Yamamoto collected cute little stuffed animals, and her roommate, Miss Morii, collected candles.'

Comparison of data produced in this experiment with that of Experiment 1 should clarify whether the contextual biasing in Experiment 1 produced an unnatural pattern of responses by helping subjects to predict the target. If the context has no reliable influence on the reading of the *kana*, then the response time patterns should replicate Experiment 1. Targets in this experiment were placed in the same syntactic position as their counterparts were in Experiment 1, although the syntactic structures of the respective sentences was not necessarily consistent. (A full set of materials is available from the first author.)

## PROCEDURE

The same procedure was followed as for Experiment 1, including the completion of post-test questionnaires by each subject. The experimental design was also the same.

## RESULTS

For Experiment 2, subject results forced us instead to set our lower limit for accuracy in the probe word identification at 85%. Five subjects, not included in the count of 20 given above, did not meet this new criterion. Again, subjects who could not read a *kanji* form that had appeared in their version of the experiment had their reading times for that form eliminated from the data. These errors accounted for 18% of the *kanji* trials overall, but all fell within the *kana* dominant condition, which had a resulting error rate of 53%.

Means were taken of the reading times for the *kanji* and *kana* forms of each stimulus across subjects and items for each experimental condition. Results for subjects are summarized in Figure 4. Reading times for each form of each item that did not fall within 2.5 standard deviations above and below the item mean were replaced with the cut-off value. Four percent of the values were replaced in this way. The final sets of reading times were analyzed in the same manner as in Experiment 1. The reading times for one *kana* dominant word, *jinmashin* 'nettle rash', were removed before analysis due to a grammatical error in the embedding sentence which occurred before the stimulus.

Contrary to our first study, the results from this experiment suggested that the script in which a word was presented had no effect on how quickly it was read by subjects, regardless of the word's orthographic dominance. There was a main effect of order by subject [ $F_1(1,17) = 6.940, p < .05$ ] and by items [ $F_2(1,17) =$

22.514,  $p < .05$ ). Orthographic dominance produced a marginal effect by subjects [ $F_1(2,34) = 2.943$ ,  $p < .1$ ] yet a main effect by items [ $F_2(1,17) = 5.828$ ,  $p < .05$ ]. By items there was also a marginal effect of orthography of presentation [ $F_2(2,34) = 2.758$ ,  $p < .1$ ], as well as a marginal interaction of orthographic dominance and order, [ $F_2(2,34) = 2.789$ ,  $p < .1$ ]. The (lack of) interaction for dominance and orthography of presentation by subjects is summarized in Figure 4.

There was no difference between the reading times in the neutral condition for *kanji* and *kana* by subjects, [ $F_1(1,17) = .712$ ,  $p > .05$ ], or by items, [ $F_2(1,17) = 1.813$ ,  $p > .05$ ]. There was also no significant difference between orthographies in the *kanji* dominant condition by subjects or items, [ $F_1(1,17) = .005$ ,  $p > .05$ ;  $F_2(1,17) = .004$ ,  $p > .05$ ], or in the *kana* dominant condition [ $F_1(1,17) = 2.040$ ,  $p > .05$ ;  $F_2(1,17) = 1.972$ ,  $p > .05$ ]. Across orthographies, *kana* dominant words were read more slowly than orthographically neutral words, which were read more slowly than *kanji* dominant words.

The results of the familiarity ratings are summarized in Table 3 below.

	<i>Kanji</i> Dominant	<i>Kana</i> Dominant	Ortho Neutral
<i>Kanji</i>	6.92	2.38	5.30
<i>Kana</i>	1.69	6.21	4.51

Table 3. Experiment 2 mean familiarity ratings for *kanji* and *kana* forms of stimuli by dominance

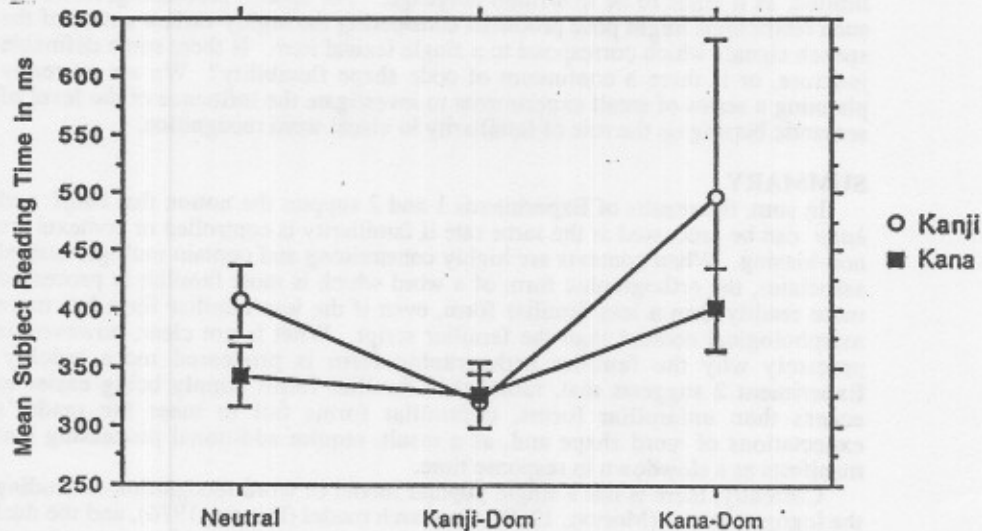


Figure 4. Mean reading time by orthography and orthographic dominance for Experiment 2, with standard error bars

## DISCUSSION

In this follow-up experiment, we found no significant difference between the reading rates for *kanji* and *kana* words in any of the individual dominance conditions. This suggests that the familiarity effects observed in Experiment 1 were due to contextual priming. In fact, the pattern of results across the two experiments suggests that *kana* forms were inhibited in the *kanji* dominant condition in Experiment 1. This might well occur if contextual priming does not only stimulate particular items in the lexicon, but also suppresses the less familiar forms. Suppose that when an unfamiliar form of the lexical item is encountered in a sentence, the processor requires additional time to deal with the unanticipated input, almost as if a word which did not suit the context had been presented. Conversely, when the context is not predictable, all forms of a lexical item are equally available. From this we may posit that the familiar orthographic form of a word is recognized more readily because it meets more of the reader's expectations--expectations which appear to be extremely specific.

This "expectation hypothesis" is consistent with the findings of Altarriba et al. (1993), who found very similar patterns for lexical targets embedded in both highly constrained and unconstrained carrier sentences using eye-tracking and naming paradigms. In contrast to our work, however, their study utilized Spanish-English bilinguals as subjects, and manipulated the language of the target word. An effect of language was found only when the sentence was strongly biased toward the target and the target was of high frequency. Since low frequency words were not affected, it is unlikely that this effect arises at the level of visual encoding. They argued that sentence context can influence expectations for upcoming words at both the semantic/conceptual level and the lexical form level.

If we are capable of accessing lexical entries in both a nonspecific and orthography specific manner, then how do we shift between the two strategies? Having very specific requirements on the shape of the code necessary for semantic access to occur would be quite efficient if the range of code shapes was limited, as it tends to be in written language. For speech processing, however, such restrictions might pose problems considering the highly variant nature of the speech signals which correspond to a single lexical item. Is there some definable juncture, or is there a continuum of code shape flexibility? We are currently planning a series of small experiments to investigate the influence of the level of semantic biasing on the role of familiarity in visual word recognition.

## SUMMARY

In sum, the results of Experiments 1 and 2 support the notion that *kanji* and *kana* can be processed at the same rate if familiarity is controlled or contexts are non-biasing. When contexts are highly constraining and contain multiple lexical associates, the orthographic form of a word which is most familiar is processed more readily than a less familiar form, even if the less familiar form has more morphological content than the familiar script. What is not clear, however, is precisely why the familiar orthographic form is processed more quickly. Experiment 2 suggests that, rather than familiar forms simply being easier to access than unfamiliar forms, unfamiliar forms fail to meet the reader's expectations of word shape and, as a result, require additional processing that manifests as a slowdown in response time.

Currently, there is not a single popular model of word recognition--including the logogen model (Morton, 1969), the search model (Forster, 1976), and the dual access model (Kleiman, 1975)--that can account for our data without a number of modifications. These models are all based on English, and assume the language of the speaker to be mono-orthographic. Since language acquisition is an innate human skill unrelated to the particular language itself, it does not seem reasonable

to posit prominent differences in the ways alphabetic, syllabic, and logographic writing systems access the lexicon. Accordingly, any model of word recognition that is truly generalizable should account equally well for input from each of these visual forms. The absence of such a universal model from the literature suggests that there is still a great deal of work to be done in the area of word recognition, and it is our hope that this study will initiate further investigation into this fascinating area of research.

## REFERENCES

- Allport, D.A. (1979) Word recognition in reading: a tutorial review. In *Processing of Visible Language Vol. 1*. (P.A. Kollers, H. Bouma, and M. Wroldstad, editors). New York: Plenum Press. 227-257.
- Altarriba, J., Kroll, J.F., Sholl, A., and Rayner, K. (1993) The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixation times and naming times. Manuscript submitted for publication.
- Aoki, C. (1990) *Hemispheric lateralization of Japanese Kanji and Kana: evidence for right hemisphere involvement in semantic processing of kanji*. Ph.D dissertation, Northeastern University.
- Besner, D. and Hildebrandt, N. (1987) Orthographic and Phonological Codes in Oral Reading of Japanese Kana. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13** (2), 335-343.
- Caplan, D., Holmes, J. M., and Marshall, J. C. (1974) Word class and hemispheric specialization. *Neuropsychologia*, **12**, 331-337.
- Day, J. (1977) Right-hemisphere language processing in normal right-handers. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 518-528.
- Ellis, H. D. and Shepard, J. W. (1974) Recognition of abstract and concrete words presented in the left and right visual fields. *Journal of Experimental Psychology*, **103** (5), 1035-1036.
- Elman, J. L., Takahashi, K., and Tohsaku, Y. (1981) Lateral asymmetries for the identification of concrete and abstract Kanji. *Neuropsychologia*, **19** (3), 407-412.
- Forster, K.I. (1976) Accessing the mental lexicon. In Garnham, A. (1985) *Psycholinguistics: Central Topics*. Routledge: London. 51.
- Goryo, K. (1987) *Yomu to Iukoto (Reading in Japanese)*. Tokyo: Tokyo University Press.
- Hatta, T. (1976) Asynchrony of lateral onset as a factor in difference in visual field. *Perceptual and Motor Skills*, **42**, 163-166.
- Hatta, T. (1977) Recognition of Japanese Kanji in the left and right visual fields. *Neuropsychologia*, **15**, 685-688.
- Hatta, T. (1978) Recognition of Japanese Kanji and Hirakana in the left and right visual field. *Japanese Psychological Research*, **20** (2), 51-59.
- Hirose, T. (1984) The effect of script frequency on semantic processing of Kanji and Kana words. *The Japanese Journal of Psychology*, **3**, 173-176.
- Hirose, T. (1985) The effects of orthographic familiarity on word recognition. *The Japanese Journal of Psychology*, **56**, 44-47.
- Hines, D. (1976) Recognition of verbs, abstract nouns and concrete nouns from the left and right visual half-fields. *Neuropsychologia*, **14**, 211-216.

- Hines, D. (1977) Differences in tachistoscopic recognition between abstract and concrete words as a function of visual half-field and frequency. *Cortex*, **13**, 66-73.
- Hirata, K. and Osaka, R. (1967) Tachistoscopic recognition of Japanese letter materials in left and right visual fields. *Psychologia*, **10**, 1-10.
- Inoue, M., Saito, H., and Nomura, Y. (1979) Psychological research on characteristics of Kanji: The effects of graphemic and phonetic processing on information extraction from Kanji. *Shinrigaku Hyoron*, **22** (2), 143-159.
- Kimura, Y. (1984) Concurrent vocal interference: Its effects on Kana and Kanji. *Quarterly Journal of Experimental Psychology*, **36A**, 117-131.
- Kleiman, G.M. (1975) Speech recoding and reading. *Journal of Verbal Learning and Verbal Behavior*, **14**, 323-339.
- McCusker, L.X., Hillinger, M.A., and Bias, R.G. (1981) Phonological recoding and reading. *Psychological Bulletin*, **81**, 217-245.
- Morton, J. (1969) Interaction of information in word recognition. *Psychological Review*, **76**, 165-178.
- Morton, J. and Sasanuma, S. (1984) Lexical access in Japanese. In *Orthographies and reading: Perspectives from cognitive psychology neuropsychology and linguistics*. (L. Henderson, editor). London: Erlbaum. 25-42.
- Ohnishi, H. and Hatta, T. (1980) Lateral differences in tachistoscopic recognition of Kanji-pairs with mixed image values. *Psychologia*, **23**, 233-239.
- Sasanuma, S., Itoh, M., Mori, K., and Kobayashi, Y. (1977) Tachistoscopic recognition of Kana and Kanji words. *Neuropsychologia*, **15**, 547-553.
- Sasanuma, S., Sakuma, N., and Tatsumi, I. (1988) Lexical access of Kana words and words in Kana. *Annual Bulletin RILP*, **22**, 117-123.
- Shanon, B. (1979) Lateralization effects in lexical decision task. *Brain and Language*, **8**, 380-387.
- Tabossi, P. (1988) Accessing Lexical Ambiguity in Different Types of Sentential Contexts. *Journal of Memory and Language*, **27**, 324-340.

## Rate Effects on German Unstressed Syllables\*

Stefanie Jannedy

jannedy@ling.ohio-state.edu

**Abstract:** German is characterized by the rhythmic alternation of strong and weak syllables. Weak syllables contain short or reduced vowels like schwa. In some instances, the unstressed weak syllable nucleus can be the only difference between words that underlyingly contain a consonant cluster. Examples in German are *Kannen* 'cans, pitchers' contrasting with *kann* 'can (V)' or *beraten* 'to advise' contrasting with *braten* 'to fry'. In some instances, in a faster rate of speech for example, weakening of the unstressed syllable nucleus is observed which can eventually result in the neutralization between such pairs of words. In slower speech, one might find an opposite effect, that is the appearance of vocalic traces between the members of an underlying consonant cluster. This transition vowel can perceptually cause a confusion in these "minimal pairs". Based on acoustic measurements, I will argue that gestural reorganization can best account for both of these rate effects found in German.


### 1. Introduction

Weakening of the unstressed syllable nucleus in German has been described and explained in terms of a phonological deletion rule (Kloeke, 1982). In recent years however, alternative explanations based on gestural reorganization have been proposed for such observations (Kohler, 1990; Browman & Goldstein, 1989, 1990a). A gestural reorganization account assumes a gradual weakening of the unstressed syllable nucleus due to overlap of adjacent consonantal gestures. According to the Gestural Score Model (Browman & Goldstein, 1989, 1990a), gestures are performed by individual articulatory subsystems. Depending on the rate of speech, the model makes two different kinds of predictions: in faster or more casual speech, articulatory gestures can overlap to a greater or lesser extent.

---

\* I am thankful to my advisor, Mary Beckman, for discussions, valuable comments and more. Keith Johnson and Sun-Ah Jun also provided much helpful advice and have taken much time to discuss issues that arose while writing this paper. Beth Hume and Julie Boland provided helpful comments on an earlier draft of this paper. Jennifer Venditti, K. Brettonel Cohen, Chip Gerfen and Bettina Migge never got tired talking about reduction processes. I also need to thank my subjects (especially Ben) who repeated the corpus over and over or patiently listened to all the stimuli. The Department of Linguistics, Ohio State University and the University of Hamburg (Germany) are thanked for their support in this research. An earlier version of this paper was presented at the 68th Annual Meeting of the Linguistic Society of America, January, 1994.

In the case of a neutralization of a contrast, a gesture, in this case the one for the unstressed vowel, is completely overlapped and therefore hidden, so that no acoustic output is generated. In theory, a second prediction is that in a slower rate of speech, the gestures for adjacent consonants in a cluster can become separated



during the transition. Depending on the degree of separation, gradually, vowel-like traces in the formant structure or even vowels of more than 20 ms in duration can appear where underlyingly not present. Phonological accounts on the other hand describe this phenomenon in terms of a categorical insertion rule (Hall, 1992).

Browman & Goldstein (1990a) provide x-ray microbeam data in support of their Gestural Score Model. In their example of the phrase *perfect memory*, the individual articulatory movements that were traced over time show that the closing gesture for the [k] in *perfect* [p<sup>h</sup>ɜfɛkt] hides the closing part of the gesture of the [t], and the closure for the gesture of the bilabial [m] in the word *memory* [memɔ:ɪ] hides the release of the [t] on the tongue tip tier. There is no acoustic output from the alveolar gesture since the gesture of the [t] is hidden by the adjacent consonantal gestures on different, independent articulatory tiers. Similarly, Munhall and Loefquist (1992) provide data that suggests gestural overlap of adjacent glottal gestures in English. They had speakers say the phrase *kiss Ted* in various speech rates and focused on the glottal aperture at the word boundary between the [s] of [kɪs] and the aspirated [t] of [t<sup>h</sup>ɛd]. In the slowest renditions, they found two distinct glottal opening gestures, in an intermediate tempo, the gestures begin to blend and the one for the [s] becomes a shoulder of the gesture for the aspirated [t<sup>h</sup>]. In the fastest tempo, the two gestures have completely blended, so that only one glottal opening gesture is observable, and the [t] acoustically has lost the aspiration because no pressure could build up. Munhall's and Loefquist's data provide evidence for the blending of gestures on identical tiers.

With respect to the German weakening phenomena, the Gestural Score Model predicts gradually decreasing vocalic durations due to various degrees of overlap, or gradually appearing and increasing vocalic durations due to the gradual separation of articulatory gestures. Phonological accounts predict either a categorical deletion or a categorical insertion. In perception, we expect gradually poorer identification scores in faster rates for the word that contains the unstressed vowel since it becomes more and more reduced. The same prediction holds in cases where vowels gradually appear where not part of the underlying gestural score. If categorical phonological rules are at work, identification will be perfect if the vowel is present in the word that underlyingly contains the vowel or if the vowel is not present in the word that underlyingly does not contain the vowel. According to the Gestural Score Model, in the case of deletion, identification of words that only contrast by the appearance of the unstressed vowel should be impossible since there will be no contrast between forms that underlyingly do not contain a vowel and the forms that underwent vowel deletion. If there is a categorical insertion, we would expect identification to be impossible too, since



there would not be a contrast between forms that underlyingly contained a vowel and the ones that underwent vowel insertion. To test these predictions, an acoustic study was performed.

### 1. Corpus and Methods

A paragraph was constructed that contained three target minimal pairs:

1. <i>Kannen</i>	[k'anən]	'cans, pitchers'
<i>kann</i>	[kan]	'can, (V)'
2. <i>geleiten</i>	[gə.l'ar.tən]	'to accompany'
<i>gleiten</i>	[gl'ar.tən]	'to slide'
3. <i>beraten</i>	[bə.ɾ'a.tən]	'to advise'
<i>braten</i>	[bɾ'a.tən]	'to fry'

Both members of each 'minimal pair' occurred within a context where the adjacent segments were identical. Six native speakers of a northern German dialect (as spoken in the south of Hamburg) read the corpus ten times each in self-selected speech rates. Speakers were instructed to produce rendition one and six at a normal rate, 2 through 5 increasingly faster relative to the previous reading and 7 through 10 slower and slower relative to the preceding reading. Duration measurements of the target words were done twice by the same person on a Kay Sonograph Spectrogram 5500-1. The measured values differed from the cross-checked values only minimally. Waveform and amplitude traces were in one display window and a wide-band spectrogram was displayed in a second window.

From each of the target words, the following measurements were taken: 1. total duration as defined from the release of the initial burst to the end of the final nasal. 2. the duration of the nasal sequence in *Kannen* and *kann*, and the duration from the onset of the second vowel to the end of the nasal in pairs two and three. 3. the VOT, from the release of the stop burst to the onset of voicing of the vowel, was taken for *Kannen* and *kann*. 4. duration from the initial burst to the end of the [a] in the first pair of words and to the end of the liquid [gl] in the second. For the *beraten* and *braten*, segmentation proved to be difficult and was cross-checked a third time with a computerized speech analysis system (Milinkovic). Previously, using the Kay, the end of the uvular fricative was determined by the dip in the amplitude trace. In C-speech, however, the end of the uvular fricative was determined by the onset of the decreased F2-bandwidth for the following [a]. The latter measurements were used in the plots of the production graphs. The appearance of a vowel in *braten* was judged by decreased F2 bandwidth right after the release of the initial voiced stop burst that then increases for the fricative. If this initial period in which we can observe an increased bandwidth was sustained for 20 ms or longer, it was judged to be containing a vowel.

For the perception test, 360 target words were spliced out of context, digitized, randomized and played back onto tape. There were also 120 filler words mixed into the randomized list of target words. Two tapes were prepared, each containing the same items but in a different order. 24 equally long blocks of 12



stimuli were recorded, each stimulus was played twice with an inter stimulus interval of two seconds. 25 native speakers of various German dialects participated in the forced choice identification perception test.

### 3. Results

The results of the production for *Kannen* and *kann* are shown in Fig.1. The durations of /ka/ of the monosyllable *kann* 'can, (V)' or of the disyllable *Kannen* 'cans, pitchers', are plotted on the x-axis as a measure of the speech rate. The target segments, that is for *kann* the /n/ (hollow circles) and for *Kannen* the nasal sequence /nən/ (filled circles) are plotted on the y-axis. The regression function indicates that as the rate of speech becomes faster and faster, the duration of the nasal sequence of *kann* also gradually shortens, but not as much as the one for *Kannen*.

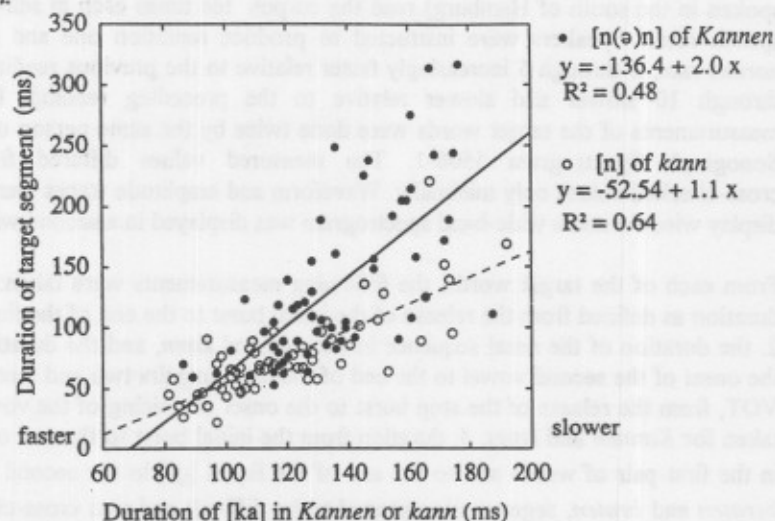


Fig.1: The duration of the nasal sequence of *Kannen* and *kann* is plotted against the remainder [ka] of *Kannen* or *kann* respectively.

The two regression functions for *Kannen* and *kann* cross as the rate of speech becomes faster. They show that the values for the duration of the [nən] and [n] incline toward each other, indicating that the duration of the nasal sequences in the monosyllable and the disyllable are of similar or equal duration in the faster rates but quite distinct in slower renditions. Yet, there is no categorical shift from

presence to absence of the vowel in the nasal sequence /nən/ of *Kannen* since we do not find two clearly separated clouds of data. There were in total seven vocalic appearances of at least 20 ms in duration in all the tokens of *Kannen*. The spectrograms in Fig.2 show tokens of *Kannen* from continuous speech rates, uttered by the same female speaker. The token on the left was produced in rate 9 (slow rate), the one in the middle in rate 8 and the one on the right in rate 7, slightly slower than normal. Whereas there are very clear vocalic traces in the spectrogram on the left, the vowel is already shorter in the middle display and eventually totally disappears, as in the spectrogram to the right. The underlying vowel only appears in relatively slow and carefully articulated speech of two speakers in this study. However, the production data does not show any discontinuities but rather a very gradual shortening, as seen in the figure.

A paired t-test of the durations of the nasal sequences in *Kannen* vs. *kann* showed that the means of the samples were significantly different at the  $p \leq 0.01$  level ( $t = -8.56, p < .001$ ). However, the patterning of the individual datapoints supports the notion that the disappearance of the vowel, or the shortening of the [n]-sequence in *Kannen* is a gradual process rather than a categorical one.

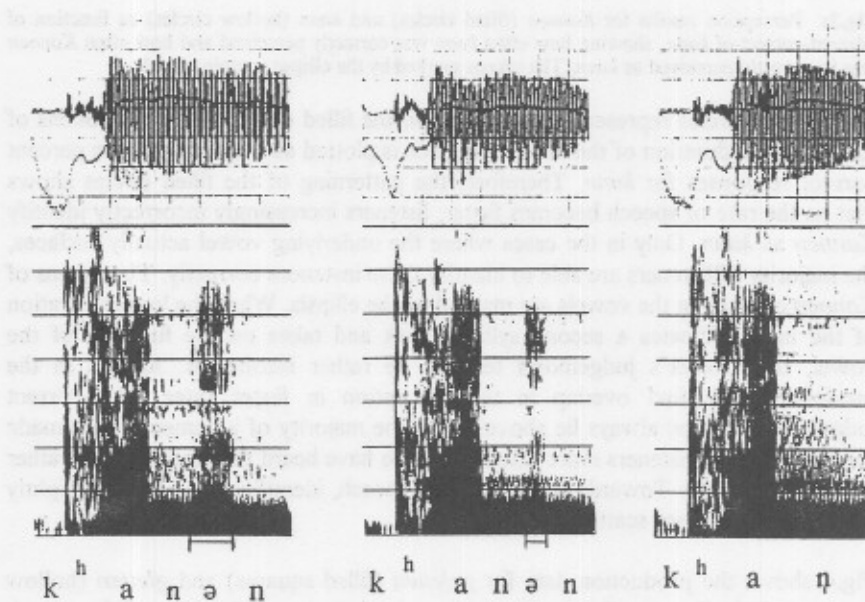
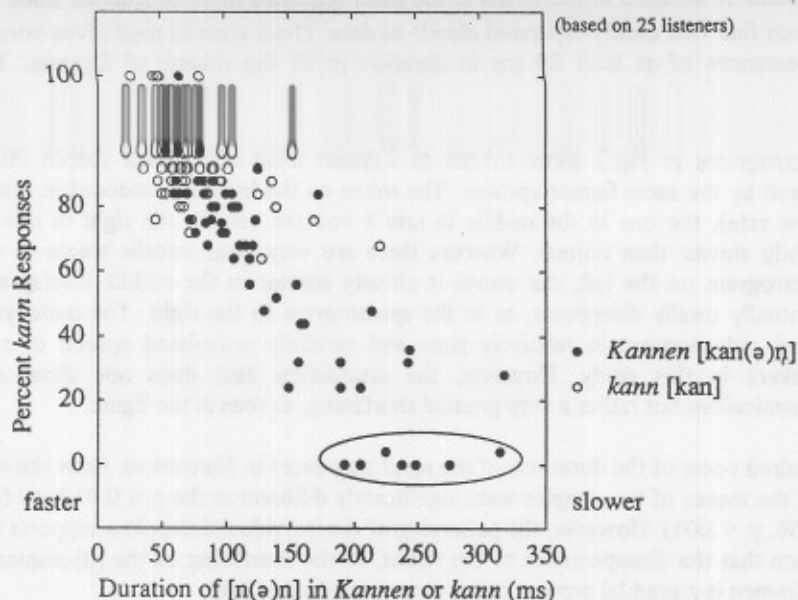


Fig.2: Spectrograms and waveform of tokens of *Kannen*, as produced by a female speaker in the rates 9, 8 and 7 from left to right.

The perception data in Fig.3 is based on the judgments of 25 listeners. The duration of the target sequences [n] and [nən] are plotted on the x-axis whereas the percent *kann* responses are plotted on the y-axis.



**Fig.3:** Perception results for *Kannen* (filled circles) and *kann* (hollow circles) as function of percent-correct of *kann*, showing how often *kann* was correctly perceived and how often *Kannen* was incorrectly perceived as *kann*. The tokens marked by the ellipse contain vowels.

The hollow circles represent tokens of *kann*, the filled circles stand for tokens of *Kannen*. The duration of the target sequence is plotted as a function of the percent correct responses for *kann*. Therefore, the patterning of the filled circles shows that as the rate of speech becomes faster, listeners increasingly incorrectly identify *Kannen* as *kann*. Only in the cases where the underlying vowel actually surfaces, the majority of listeners are able to identify these instances correctly. The tokens of *Kannen* containing the vowels are marked by the ellipsis. When the longer duration of the nasal indicates a second syllabic peak and takes on the function of the vowel, the listener's judgements tend to be rather incoherent. Just as in the production, we find overlap in the perception in faster rates. The correct judgments for *kann* always lie above 60%. The majority of judgments were made for *kann*, that is, listeners more often judged to have heard the monosyllable rather than the disyllable. Towards slower rates of speech, identification becomes slightly worse and values are scattered.

Fig.4 shows the production data for *geleiten* (filled squares) and *gleiten* (hollow squares). In both cases the target sequence is plotted as a function of the speech rate, manifested by the remainder of the word, which is [atən]. The values are non-overlapping and throughout all speech rates, a difference between the two target portions of *geleiten* and *gleiten* is retained. A paired t-test shows that [gəl] and [gl] are significantly different at the .01 level ( $t = 1.51$ ,  $p = 0.0$ ).

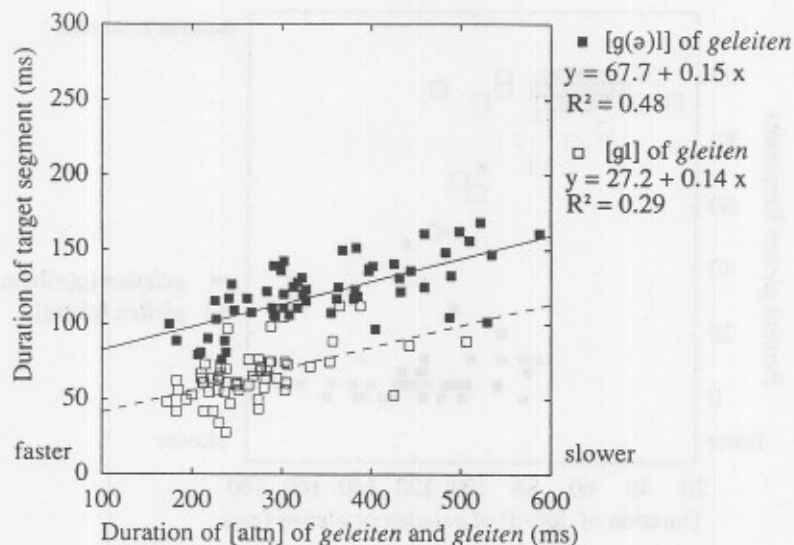


Fig.4: The duration of the remainder is plotted against the target segments. Filled squares show cases of *gezeiten* and hollow squares represent instances of *gleiten*.

The regression functions indicate that as the rate of speech becomes faster, the values for the duration of the target sequences do not incline toward each other. All tokens of *gezeiten* contained a vowel in all rates. If there was a categorical deletion, we would expect two distinct clouds of data for *gezeiten* in production. This however is not the case; there is only one cloud of data for the underlyingly three syllable word. Also, the values for *gleiten* versus *gezeiten* are fairly distinct, predicting good identification in the perception test. We can observe, however that as the rate of speech becomes faster, the duration of the target sequences [gəl] and [gl] becomes shorter in both cases, indicating that the faster rate of speech influences the duration of the target sequences.

Fig.5 shows the perception data for *gezeiten* and *gleiten*. The duration of the target segments [gəl] (filled squares) and [gl] (hollow squares) is plotted as a function of the percent *gleiten* responses. Although identification is generally very good and we find two very distinct clouds of data which shows that the perception in these cases is rather categorical, there are a few cases of *gezeiten* as well as *gleiten* that had a tendency to be misjudged more often. Apart from these few tokens, the majority of the words were correctly identified well above chance level. It is not clear what led listeners to make the few incorrect judgements.

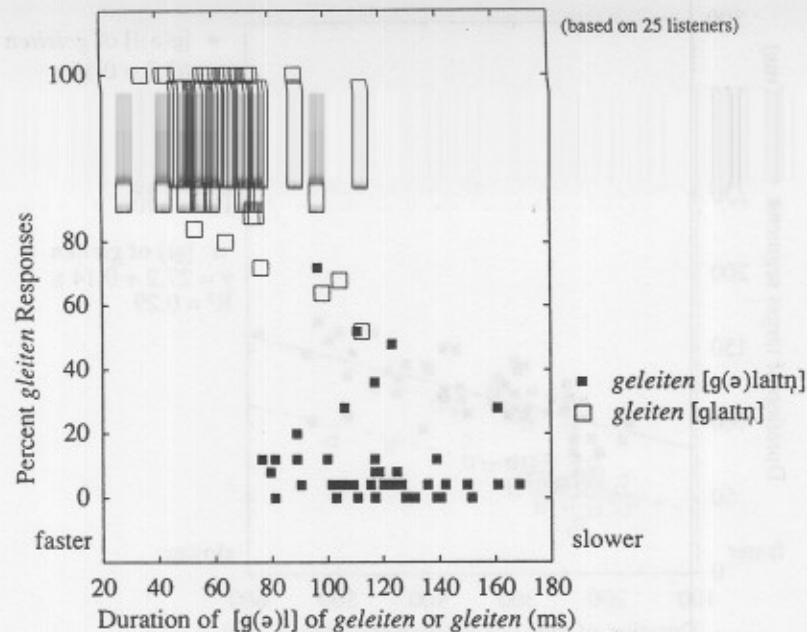
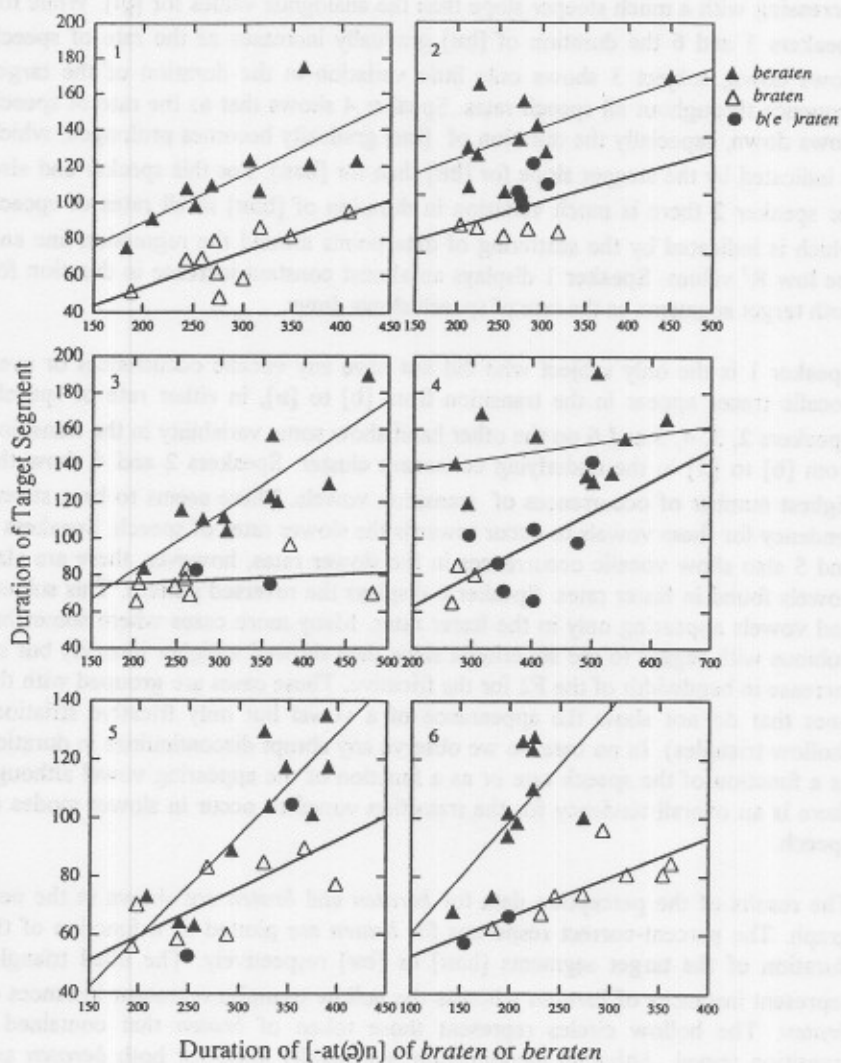


Fig.5: Perception data for *geleiten/gleiten*. The duration of the target sequence is plotted against the percent *gleiten* responses.

Fig.6 shows the production data for *beraten* and *braten*, for the six speakers. The graphs are presented individually to show the general tendencies for each speaker. These individual graphs show that speakers have different strategies and that one speaker's fastest rate can be another speaker's normal rate. The duration of the target segment [bæʁ] or [bɪʁ] is plotted on the y-axis and the remainder of the token, that is [atŋ] of either *beraten* or *braten*, is plotted on the x-axis. A paired t-test showed that these two target segments were significantly different at the .01 level ( $t = 1.31$ ,  $p < .001$ ). The filled triangles are tokens of the trisyllabic word *beraten* and the and hollow triangles symbolize tokens of *braten*. The filled circles stand for cases of *braten* where clear vowels of at least 20 ms of duration were found between the release of the voiced bilabial stop [b] and before the increase of the bandwidth of F2 during the uvular/velar fricative (due to decoupling of the front and the back cavity in the vocal tract). The values for the slopes and the intercepts of *beraten* and *braten* respectively are as follows for graphs one through six:

1.  $y = 36.8 + 0.3 x$ ,  $R^2 = 0.53$ ;  $y = 14.7 + 0.2 x$ ,  $R^2 = 0.56$ ;
2.  $y = 82.4 + 0.1 x$ ,  $R^2 = 0.3$ ;  $y = 52.2 + 0.1 x$ ,  $R^2 = 0.13$ ;
3.  $y = 20.3 + 0.3 x$ ,  $R^2 = 0.71$ ;  $y = 71.1 + 0.1 x$ ,  $R^2 = 0.4$ ;
4.  $y = 128.2 + 0.1 x$ ,  $R^2 = 0.7$ ;  $y = 27.8 + 0.2 x$ ,  $R^2 = 0.37$ ;
5.  $y = -10.1 + 0.4 x$ ,  $R^2 = 0.72$ ;  $y = 26.7 + 0.2 x$ ,  $R^2 = 0.46$ ;
6.  $y = 22.0 + 0.4 x$ ,  $R^2 = 0.38$ ;  $y = 38.8 + 0.1 x$ ,  $R^2 = 0.69$ .



**Fig.6:** Production graphs for all six speakers for *beraten* (filled triangles) and *braten* (hollow triangles). The filled circles indicate the token where vowel-like segments appeared in the transition between [b] and [ɪʁ] in the *braten* cases. On the x-axis the duration of the remainder is displayed, and on the y-axis the duration of the target segment. The circles are included in the regression calculation.

There is not one consistent pattern for all six speakers in production. For speakers 3, 5, and 6, the regression functions incline towards each other in faster rates of speech, showing that as the rate of speech becomes slower, the values for the

target segments [bɔɪ] and [bɪ] diverge, with the values for [bɔɪ] gradually increasing with a much steeper slope than the analogous values for [bɪ]. While for speakers 5 and 6 the duration of [bɪ] gradually increases as the rate of speech

slows down, subject 3 shows only little variation in the duration of the target sequence throughout all speech rates. Speaker 4 shows that as the rate of speech slows down, especially the duration of [bɪ] gradually becomes prolonged, which is indicated by the steeper slope for [bɪ] than for [bɔɪ]. For this speaker and also for speaker 2 there is much variation in duration of [bɔɪ] in all rates of speech which is indicated by the scattering of data points around the regression line and the low  $R^2$  values. Speaker 1 displays an almost constant increase in duration for both target segments as the rate of speech slows down.

Speaker 1 is the only subject who did not have any vocalic occurrences or even vocalic traces appear in the transition from [b] to [ɪ], in either rate of speech. Speakers 2, 3, 4, 5 and 6 on the other hand show some variability in the transition from [b] to [ɪ] in the underlying consonant cluster. Speakers 2 and 4 show the highest number of occurrences of transition vowels. There seems to be a strong tendency for these vowels to occur towards the slower rates of speech. Speakers 3 and 5 also show vocalic occurrences in the slower rates, however, there are also vowels found in faster rates. Speaker 6 displays the reversed pattern. This subject had vowels appearing only in the faster rates. Many more cases where somewhat dubious with regard to the set criteria since they showed a higher intensity but no increase in bandwidth of the F2 for the fricative. These cases are grouped with the ones that do not show the appearance of a vowel but only fricative striations (hollow triangles). In no case do we observe any abrupt discontinuities in duration as a function of the speech rate or as a function of the appearing vowel although there is an overall tendency for the transition vowel to occur in slower modes of speech.

The results of the perception data for *beraten* and *braten* are shown in the next graph. The percent-correct responses for *braten* are plotted as a function of the duration of the target segments [bɔɪ] or [bɪ] respectively. The filled triangles represent instances of *beraten* whereas the hollow triangles represent instances of *braten*. The hollow circles represent those token of *braten* that contained a transition vowel. Although identification is generally good for both *beraten* and *braten*, more slower tokens of *braten* that contained a vowel were repeatedly misjudged as instances of *beraten*. This possibly indicates that listeners did not take speech rate differences into consideration.

However, there are also slower tokens of *braten* that did not contain a vowel and that were repeatedly misjudged. It remains unclear what cues listeners used to judge these tokens of *braten* as *beraten*. The identification rate for *braten* scatters between perfect (100%) and chance-level (50%). Two outlying tokens of *braten*, both containing a vowel, were often misjudged as *beraten*. However, there are



cases where we find a transition vowel but that were identified very well as *braten*.

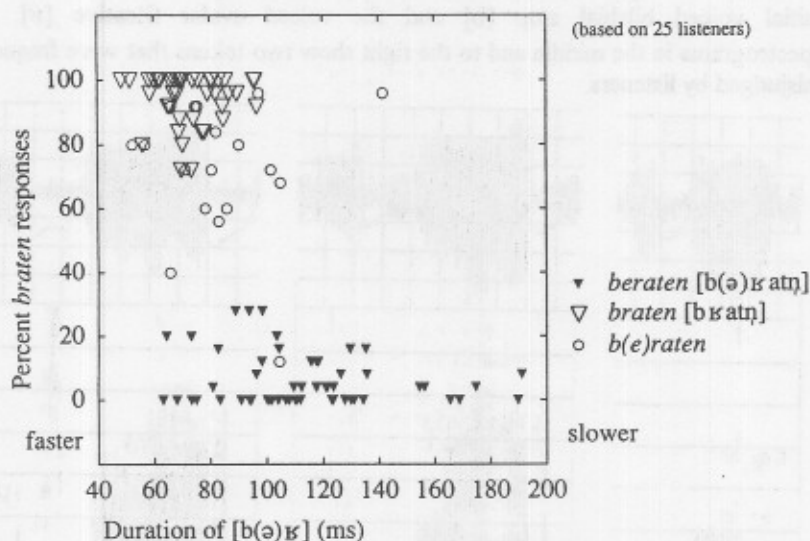


Fig.7: The duration of the target sequence is plotted against the percent *braten* responses. The hollow circles indicate which tokens of *braten* contained vowels.

Throughout all speech rates, identification of *beraten* was very good. *Beraten* was correctly perceived in at least 70% of the cases. There is a slight tendency for identification to get better the more the speech slows down, that is, the longer the target sequence becomes.

Fig.8 shows spectrograms of tokens of *gleiten* as well as *braten* to illustrate potential differences between tokens that were correctly perceived and those that listeners had trouble identifying correctly. All spectrograms in the upper panel were produced by the same male speaker. The spectrogram in the upper left panel shows an instance of *gleiten* (rate 5) that was correctly perceived by 24 out of 25 listeners. The one in the middle (rate 7) and on the right (rate 9) show two tokens of *gleiten* that were repeatedly misperceived as *geleiten* (in over 50% of the cases). Here, a sharp and sudden rise in intensity is reflected in the amplitude tracing. In the correctly perceived token, the amplitude rises into the intensity plateau of the following diphthong. In the mostly misperceived cases, the amplitude raises with a steeper slope and has already reached its maximum value during the transition into the vowel. Whether the difference in amplitude causes a difference in how the token is perceived is not clear at this point. Based on Price's (1980) experiments, however, amplitude was not a decisive factor. In the second row, we see three spectrograms of tokens of *braten* that were produced in different speech rates by the same female speaker. The spectrogram on the left

shows a token produced in a faster rate than normal that was always correctly perceived. As expected, we do not find any vocalic segment between the word initial voiced bilabial stop [b] and the voiced uvular fricative [ʁ]. The spectrograms in the middle and to the right show two tokens that were frequently misjudged by listeners.

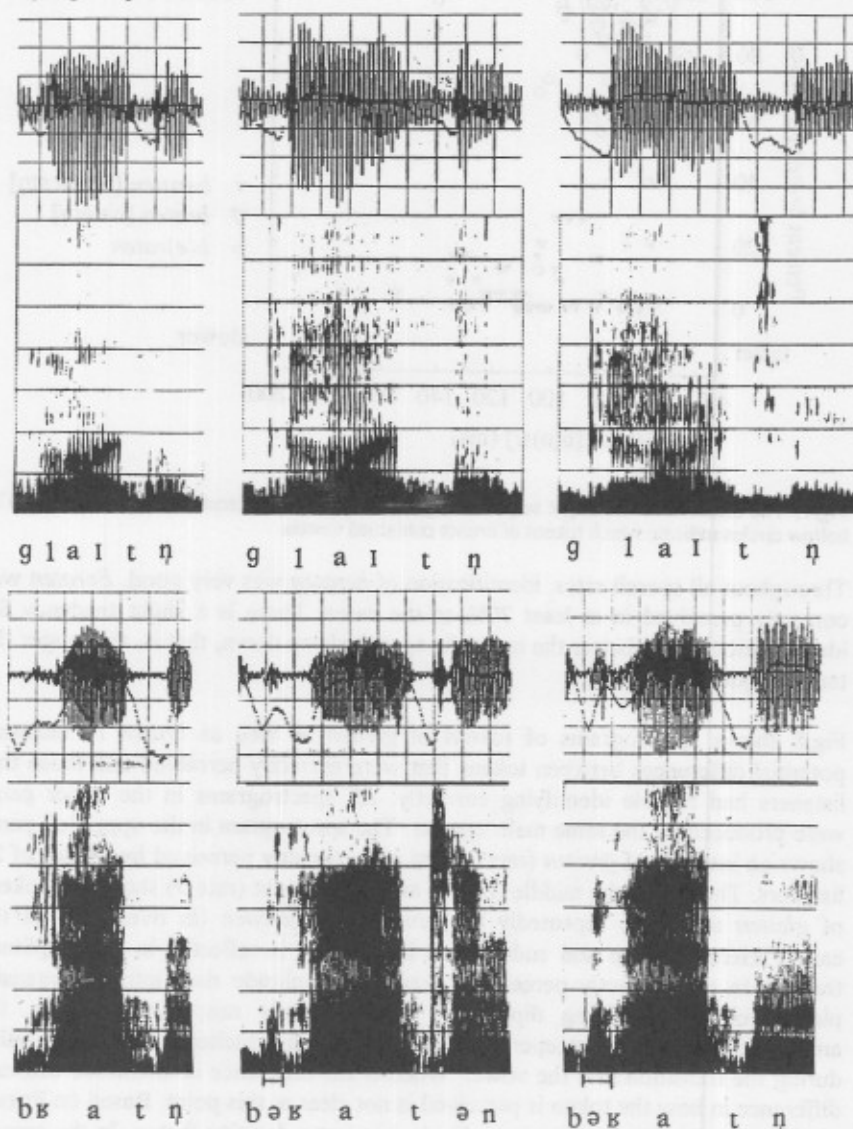


Fig. 8: Upper panel: Spectrograms of three tokens of *gleiten* (produced by the same male speaker). Left: rate 5; middle: rate 7; right: rate 9. Lower panel: Three tokens of *braten* (produced by the same female speaker). Left: rate 3; middle: rate 9; right: rate 8. In each panel, the tokens in the middle and to the right were misperceived most often.

The token in the middle was produced in a fairly slow rate (9) and the one on the right in rate (8). However, since the speakers were asked to read the corpus in self-selected speech rates, there is no absolute measure for the rate of speech. Therefore, rendition 8 can be slower than rendition 9 although it should be faster. In the second and third spectrogram, we do find vocalic traces between the initial cluster consonants. This finding is reflected in the waveform and the amplitude tracing as well. This transition vowel appears even stronger in the third display which was uttered slightly slower than the second one.

For a comparison, spectrograms for *geleiten* and for *beraten* (both in rate 1), produced by the same speakers as in the previous figure, are given in Fig.9. In the spectrogram on the left, we see clear vowel formants for the unstressed vowel, including a velar pinch, typical for a consonantal constriction in the velar region of the vocal tract. There is a sharp rise in the intensity level from the release burst of the [g] to a plateau at the unstressed vowel which is maintained throughout the [l] and the following diphthong.

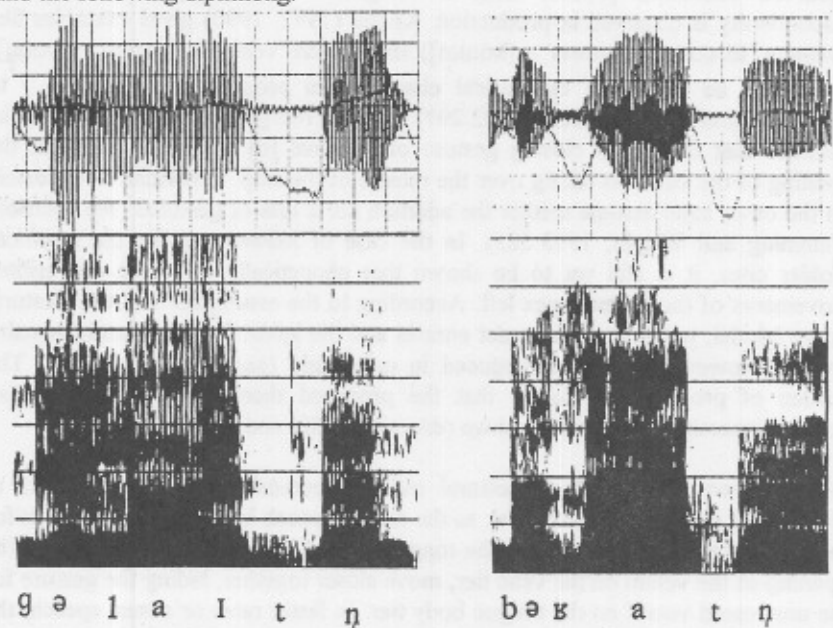


Fig.9: Spectrogram, waveform and amplitude tracings of *geleiten* (left) and *beraten* (right), produced in a normal rate by the same male and the same female speaker that produced the tokens in Fig.8.

On the right we find a clear vowel that is followed by frication with an energy concentration in the lower frequency area, typical for back fricatives. The waveform as well as the amplitude trace also clearly show the presence of the vowel before the voiced fricative. Compared to the spectrograms in Fig.8, the duration of the vowel is longer here. However, apparently, even the shorter

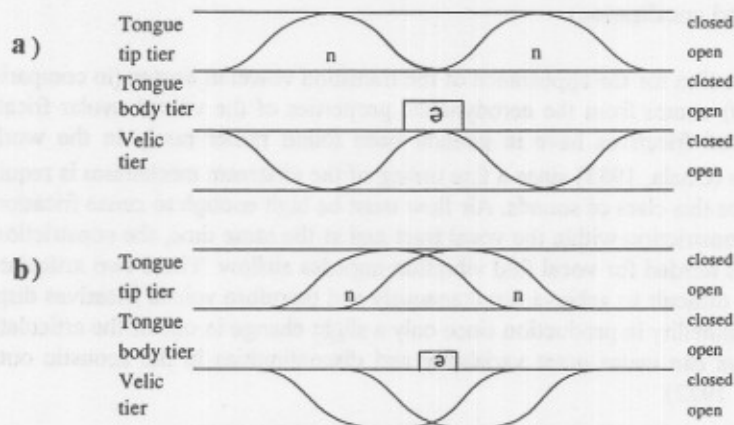
duration of the vowel is sufficient to cause a perceptual confusion for the listeners. The seemingly longer duration of the [l] in *gleiten* in the misperceived tokens might have been interpreted as a syllabic peak. It remains unclear what parameters

(amplitude, duration) listeners use. Even though the acoustic characteristics of *gleiten* vs. *geleiten* and *beraten* vs. *braten* show differences, the similarities are strong enough to cause perceptual confusion. To explore the specific perceptual cues listeners use is beyond the scope of this study.

#### 4. Discussion

Even though Kohler (1992), in a critical commentary to the Gestural Score Model, argues that the mechanism of gestural overlap alone cannot explain connected speech processes as found in cases analogous to *Kannen*, the production data presented here for *Kannen* strongly favor a gestural overlap account of the observed reduction process over a categorical deletion process since no discontinuity is observed in production. Kohler (1992; 1990) gives examples like *kommen* 'to come' (/komen/ > [kɔmm]) and *Wagen* 'vehicle' (/vagen/ > [vɑŋ]) for which he claims "a categorical change from progressive overlap [...] to reorganization" of gestures (1992:207). In his reorganization explanation he assumes that the apical closing gesture of the final [n] is deleted and that the opening of the velum is taking over the release of the stop. Browman & Goldstein on the other hand assume neither the addition nor a loss of gestures (see Johnson, Flemming and Wright, 1993:525). In the case of *Kannen*, as for the instances Kohler cites, it is still yet to be shown that phonetically there are no residual movements of the tongue apex left. According to the assumptions of the Gestural Score Model, gestures are abstract entities and the apical closing gesture remains present, however, somewhat reduced in magnitude (see Kingston, 1992). The burden of proof is on Kohler that the proposed theoretical gestural overlap account inaccurately explains schwa reduction in this and analogous cases.

Fig.10 shows a hypothetical gestural overlap account, based on Browman & Goldstein's Gestural Score Model: as the rate of speech increases, the gestures for the alveolar nasals, produced on the tongue tip tier, as well as the gesture for the opening of the velum on the velic tier, move closer together, hiding the gesture for the unstressed vowel on the tongue body tier. In faster rates or casual speech, the gestures of the alveolar nasals and the velic gestures have completely overlapped the gesture for the unstressed vowel. The vocalic gesture is hidden behind the consonantal gestures and not deleted, as has been claimed in phonological accounts.



**Fig.10:** Hypothetical Gestural Score showing in a) two alveolar gestures on the tongue tip tier for the nasals and the unstressed vowel on the tongue body tier. In b) a situation where the gestures for the nasals overlap and hide the vocalic gesture, but where some durational difference in the nasal sequence vs a single nasal is still preserved.

Various examples cited in the literature for English and other languages (see Beckman, this volume) show cases where articulatory gestures are overlapped and become hidden (Browman & Goldstein, 1989, 1990a; Munhall & Loefquist, 1992; Jun & Beckman, 1993) so that their acoustic output is much reduced which then can potentially result in the total loss of the perceptability. Also, Browman & Goldstein (1990b:314/315) showed that the identification function in synthesized stimuli on a continuum of gestural overlap between the gestures for the [b] and the rhotic (American) [ɹ] from *bray* to *beret* (generated from calculations of 'the task dynamic and vocal tract models') shifted at 0 ms overlap (perfect alignment) for four out of six listeners. Price (1980) provides similar results by modeling gestural overlap with varying the duration of [l] in synthesized stimuli of *plight* and *polite*.

No cases cited in the literature give examples or show evidence for the gradual appearance of a vocalic segment due to gestural separation even though such cases are not ruled out by Browman & Goldstein (1990b:318; 1992b:53) e.g. as a possible source for sound change. The evidence from (especially slower renditions of) *braten* favor an account of a gradual separation of articulatory gestures that results in the gradual appearance of vowel-like traces in the formant structure or even a vowel. The acoustic output during the transitions from one articulatory gesture to the next one is then misperceived. In contrast to the [l] in *plight* (Price, 1980) or the [ɹ] in *bray* (Browman & Goldstein, 1990b) where sonorant liquids were lengthened to the point that the duration gave listeners an increased percept of sonority and hence a syllabic peak (both these sounds have the property of being able to become syllabic), the German voiced uvular fricative does not seem to have

the property of becoming syllabic with an increased duration (see Kohler, 1991 on German [ʁ] vocalization).

An explanation for the appearance of the transition vowel in *braten* (in comparison to *gleiten*) comes from the aerodynamic properties of the voiced uvular fricative [ʁ]. Voiced fricatives have in general been found rather rarely in the world's languages (Ohala, 1983) since a fine tuning of the airstream mechanism is required to produce this class of sounds. Air flow must be high enough to cause friction at a given constriction within the vocal tract and at the same time, the constriction at the glottis needed for vocal fold vibration impedes airflow. These two articulatory goals are difficult to achieve simultaneously and therefore voiced fricatives display greater instability in production since only a slight change in one of the articulatory parameters can cause great variability and discontinuities in the acoustic output (Stevens, 1972).

The variability in production for *braten* can thus be explained as follows: right after the release of the bilabial occlusion for [b], there will be stronger airflow. After the air pressure has leveled out (within approximately 3 to 5 glottal pulses), the voicing sets on and the uvular constriction for the fricative can be formed again. In these cases where the articulatory gestures are misaligned, vowel-like formant structures are produced. Vowel durations between 30 and about 60 ms were found where actually no vowel is underlyingly present. In Fig.11, the appearance of this transition vowel is explained by means of Browman & Goldstein's Gestural Score Model: as the closure of the [b] on the lip tier is released, the gesture for the uvular fricative on the tongue body tier has already set on and (almost) come to a closure on the tongue body tier. There is no transition vowel produced here.

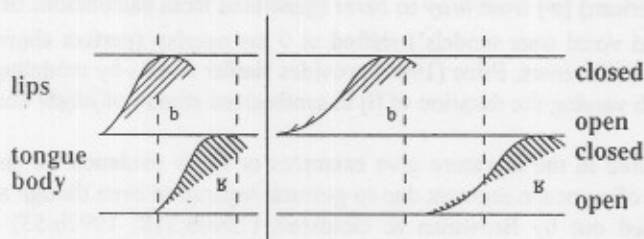


Fig.11: Hypothetical Gestural Score for the sequence [bʁ] in *braten*. On the left, the gestures are phased so that no vocalic output will be generated since the [ʁ] has already come to a closure. On the right, the gesture for [ʁ] only sets on after the gesture of the [b] has been released.

On the other hand, in the right panel, as the lips open after the bilabial closure for the [b], the tongue body has barely begun to rise, and the separation of gestures in time creates a transition vowel. Therefore, gestural separation appears to be responsible for generating an unintended vocalic segment. Even when a transition vowel appears, there will be no gesture for the vowel since it is not part of the

gestural score for [bɤ].

This misalignment of gestures and the instability of the articulation of the voiced uvular fricative is not an aerodynamic artifact since for example speaker 1 did not show any cases where formant structures in the acoustic output were visible. Nor does it only occur in slower rates of speech since speaker 6 showed a reversed pattern. The gestural separation and the resulting transition vowels are a result of the phasing of the release of the bilabial occlusion and the onset of constriction and timing of the laryngeal gesture for [ɤ]. Speakers might have some intuitive knowledge based on their experience with language about how much variation is allowed in production to still match the output goal (see Hawkins, 1992:57). Even though clear vowels appeared in the acoustic signal, the listener can apparently to some degree compensate for this and still perceive these tokens as instances of *braten*. The listener might have taken the durational difference of the vowel into account for which he or she has some kind of expectation. However, this can only be hypothesized at this point. In no case of *beraten* the schwa was reduced to such an extent that the acoustic output was completely hidden. This might be explained along the same lines as the appearance of vocalic traces and vowels in *braten*.

During the repositioning of the articulators from one articulatory gesture to the next, the tongue moves through a "neutral position" (Browman & Goldstein, 1992a:55; also see Barry, 1992 for a discussion) and the transition vowel will be generated. Phonological accounts treat the appearance of a vowel as a categorical insertion (Hall, 1992). However, temporal or syntagmatic coordination (Browman & Goldstein, 1992b) and therefore, gestural separation and the misalignment of articulatory gestures can be modeled by the Gestural Score Model. It is crucial though that the transition vowel found in some cases of *braten* does not have a target (Browman & Goldstein, 1992a) since it is not part of the underlying form.

Based on the findings of gestural overlap in faster rates of speech as exemplified in the case of *Kannen*, the appearance of the vocalic traces in slower renditions of *braten* can be now explained without positing a categorical vowel insertion rule but by the reverse mechanism of gestural overlap, that is gestural separation. The Gestural Score Model therefore provides a unified account of gestural reorganization for the observations that are traditionally explained by processes of deletion (Kloeke, 1982) or insertion (Hall, 1992; Strauss, 1982).

According to Lindblom's (1990) 'Hyper- and Hypoarticulation' Theory, speech production varies along a continuum of hypo- and hyperarticulated speech (see also Johnson, Flemming and Wright, 1993). The speaker takes the communicative demands of the situation into account and strives for sufficient discriminability. This process is controlled by output oriented feedback mechanisms. Depending on the amount of effort a speaker puts into the production of speech, the articulatory target can be undershot. In this study, for *braten*, vowels were found not exclusively but predominantly in slower renditions of speech which will be equated

with more careful, clear or hyperarticulated speech. In this study though, it was found that carefully produced speech (hyperarticulation) can result in the appearance of vowels which in turn can mislead the listener. This of course cannot be the goal for the speaker. However, the data presented for *braten* does not argue against the theory that phonetic targets are hyperarticulated. The misaligned transitions rather than an overshoot of the phonetic target, that is the syntagmatic relationship (Kohler, 1986) between one gesture's offset and an adjacent gesture's onset, cause the appearance of unintended vocalic traces or vowels which potentially cause perceptual confusion. Interestingly, the appearance of the vowel occurred in the case of [bɛ], the sequence with an aerodynamically unstable sound. Note also that the appearance of the vowel was not perfectly correlated with the rate of speech or the care of production. Previous studies did not specifically look at cases that display aerodynamic instabilities of the consonantal gestures involved forming the consonant cluster. To verify that this is a necessary prerequisite for non-underlying vowels to appear, more evidence is needed. It is conceivable also that the obstruent status of the [ɛ] influences the appearance of the transition vowel. This, however, needs more exploration.

## 5. Conclusion

Evidence for the applicability of the Gestural Score Model for German was provided. The production as well as the perception data presented for *Kannen* argue for a gradual reduction of the unstressed syllable nucleus rather than a categorical deletion process. Furthermore, for *braten*, in some instances unpredicted (i.e. non-underlying) vowels and vowel-like traces were produced. Here the opposite mechanism appears to be at play, that is a gradual separation of adjacent articulatory gestures due to a misalignment in temporal coordination. Therefore, it is concluded that the Gestural Score Model can account for both, the gradual reduction as well as the gradual appearance of vowels in German, whereas the phonological accounts needs to posit a categorical insertion as well as a categorical deletion rule.

## 6. References

- Barry, W. (1992) Comment on "Targetless" schwa: an articulatory analysis. In G. J. Docherty & D. R. Ladd, eds., *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 65-67. Cambridge: Cambridge University Press.
- Beckman, M. (this volume) When is a Syllable not a Syllable? OSU Working Papers in Linguistics, 44.
- Browman, C. P., Goldstein, L. (1989) Articulatory gestures as phonological units. *Phonology*, 6, 201-251.
- Browman, C. P., Goldstein, L. (1990a) Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman, eds., *Papers*



- in *Laboratory Phonology I: Between the Grammar and Physics of Speech*, 341-376. Cambridge: Cambridge University Press.
- Browman, C. P., Goldstein, L. (1990b) Gestural specification using dynamically-defined structures. *Journal of Phonetics*, **18**, 299-320.
- Browman, C. P., Goldstein, L. (1992a) "Targetless" schwa: an articulatory analysis. In G. J. Docherty & D. R. Ladd, eds., *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 26-56. Cambridge: Cambridge University Press.
- Browman, C. P., Goldstein, L. (1992b) Articulatory Phonology: An Overview, *Phonetica*, **49**, 155-180.
- Hall, T. A. (1992) Syllable Structure and Syllable-related Processes in German, *Linguistische Arbeiten 276*. Tuebingen: Max Niemeyer Verlag.
- Hawkins, S. (1992) Comment on "Targetless" schwa: an articulatory analysis. In G. J. Docherty & D. R. Ladd, eds., *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 56-59. Cambridge: Cambridge University Press.
- Johnson, K., Flemming, E. & Wright, R. (1993) The hyperspace effect: Phonetic targets are hyperarticulated, *Language*, **69**, 505-528.
- Jun, S.-A., Beckman, M. E. (1993) A gestural-overlap analysis of vowel devoicing in Japanese and Korean. Paper presented at the 1993 Annual Meeting of the Linguistic Society of America, 7-10 January 1993, Los Angeles, CA, USA.
- Kingston, J. (1992) Comment on "Targetless" schwa: an articulatory analysis. In G. J. Docherty & D. R. Ladd, eds., *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 60-65. Cambridge: Cambridge University Press.
- Kloeke, W. v. L. (1982) Deutsche Phonology und Morphology. Merkmale und Markiertheit. *Linguistische Arbeiten 117*. Tuebingen: Max Niemeyer Verlag.
- Kohler, K. J. (1986) Invariance and Variability in Speech Timing: From Utterance to Segment in German. In Perkell, J. S. & Klatt, D. H., eds., *Invariance and Variability in Speech Processes*, 268-289.
- Kohler, K. J. (1990) Segmental Reduction in Connected Speech in German: Phonological Facts and phonetic Explanations. In W. J. Hardcastle & A. Marchal, eds., *Speech Production and Speech Modeling*, 69-92. Amsterdam: Kluwer.
- Kohler, K. J. (1991) Synthesis of German /r/ in Text-to-Speech. In *Proceedings of the XIIIth International Congress of the Phonetic Sciences*, **3**, 490-493. Aix-en-Provence, France, 1991
- Kohler, K. J. (1992) Gestural Reorganization in Connected Speech: A Functional Viewpoint on 'Articulatory Phonology'. *Phonetica*, **49**, 205-211.
- Lindblom, B. (1990) Explaining Phonetic Variation: A Sketch in the H&H Theory. In W. J. Hardcastle & A. Marchal, eds., *Speech Production and Speech Modeling*, 403-439. Amsterdam: Kluwer.
- Munhall, K., Loefquist, A. (1992) Gestural aggregation in speech: laryngeal gestures. *Journal of Phonetics*, **20**, 111-126.
- Ohala, J. J. (1983) The Origin of Sound Patterns in Vocal Tract Constraints. In P. F. MacNeilage, ed., *The Production of Speech*, ch. 9, 189-216.

Price, P. J. (1980) Sorority and Syllabicity: Acoustic Correlates of Perception, *Phonetica*, 37, 327-343.

Stevens, K. N. (1972) The Quantal Nature of Speech: Evidence from

Articulatory-Acoustic Data. In *Human Communication: A Unified View*, Chapter 3, 51-66.

Strauss, S. L. (1982) Lexicalist Phonology of English and German. Dordrecht: Foris.

**Asymmetry of prosodic effects on the glottal gesture in Korean\***

Sun-Ah Jun

sjun@ling.ohio-state.edu

**Abstract :** Many languages have different allophones for voiced or voiceless stops depending on position within the word or the phrase (Keating et al. 1983). However, such effects are not always symmetrical. In this paper, I examined the voicing of the word final lenis stop when it comes at the end of the Accentual Phrase. By contrast to the word initial lenis stop, which is almost always voiceless at the beginning of the Accentual Phrase (Jun 1990a,b, 1993), the word final lenis stop was voiced at the resyllabified Accentual Phrase initial position. The data showed that the voicing of lenis stop depends on its duration relative to the following vowel and this duration was determined by its position relative to the prosodic contexts. Therefore, I proposed that the Lenis Stop Voicing rule in Korean is not a phonological rule, but is a byproduct of some other effect of prosodic position on the gestural amplitude and overlapping, thus producing a continuum of voicing. To distinguish the different duration pattern of the lenis stop, thus the different voicing pattern of the lenis stop, I suggested different prosodic representations utilizing the coda/onset information.


**1. Introduction**

It is well established that prosody conditions segmental and suprasegmental features. For example, in English, the 'gestural magnitude' of /h/ is weakened in word medial position or in deaccented words so that overall amplitude is smaller and energy is more concentrated in the first harmonic (Pierrehumbert and Talkin 1992). Also, segments are found to be lengthened at the edge of a phrase (e.g. Oller 1973; Beckman and Edwards 1990). As shown in Keating et al.'s (1983) survey of phonetic studies, many languages have different allophones for voiced or voiceless stops depending on position within the word or the phrase. However, such effects are not always symmetrical. For example, in German, voiced stops often become voiceless word initially as well as word finally, but this causes neutralization only word finally, where the contrasting voiceless stop is not aspirated.

Korean also has such prosodically conditioned strengthenings and weakenings of laryngeal features, and asymmetries between word final and non-final position. In the initial position of a word in isolation, there is a three way contrast among aspirated, tense, and lenis voiceless obstruents. In word medial position, the stops are weakened so that the aspirated stops are less aspirated (Jun 1990a, 1993) and the lenis stops are voiced intervocalically. In final position, the

\* The revised version of this paper will appear in *Papers in Laboratory Phonology IV*. B. Connell and A. Arvaniti, eds., Cambridge University Press, England. I would like to thank Mary Beckman, Michel Jackson, and Janet Pierrehumbert for their comments and suggestions. The work reported in this paper was supported by the NSF under Grant No. IRI-8858109 to Mary E. Beckman.

distinction is neutralized completely to an unreleased lenis stop. Kagaya (1974) shows the laryngeal configuration of the three types of obstruents; the aspirated stop has a large glottal opening gesture which peaks around the oral release, while



the tense or fortis stop has a much smaller opening and is closed even before the oral release. This is true both in initial and medial position. However, the glottal pattern of the lenis stop is more variable: in initial position a lenis stop has a large glottal opening, like that of the aspirated stop, but timed differently so that voicing starts soon after oral release. In word medial intervocalic position, it shows no glottal opening and the closure duration is very short enhancing the percept of voicing.

The laryngeal adjustment of coda obstruents, utterance finally and before other obstruents was examined in Sawashima et al. (1980). They show that the laryngeal gesture of a word final obstruent at the end of a sentence has a small glottal opening, which begins at or slightly after the oral closure and remains open for about 80-100 ms. (The laryngeal feature of the syllable final obstruent followed by other obstruent appears to be assimilated to that of the following obstruent. No fiberoptic data is available for the laryngeal gesture of the word final obstruents followed by a vowel initiated word.) That is, a coda obstruent is neutralized to an unreleased voiceless lenis stop.

Then, what is the domain of coda neutralization? The examples in (1) show that the domain is a stem plus a case marker, the Prosodic Word (Kang 1992). (Here, a dot refers to a syllable boundary.)

- (1) a. /tʃip/ => [tʃip<sup>o</sup>] 'a house'  
       /tʃip<sup>h</sup>/ => [tʃip<sup>o</sup>] 'straw'  
       /tapʃi/ => [tap<sup>o</sup>.tʃi] 'an answer sheet'
- b. /tʃip-in/ => [tʃi.bin]<sup>1</sup> 'a house-TOP'  
       /tʃip<sup>h</sup>-i/ => [tʃi.p<sup>hi</sup>] 'straw-NOM'
- c. /tʃip+/ilim/ 'name' => [tʃip<sup>o</sup>.i.rim] or [tʃi.bi.rim] 'the name of a house'  
       /tʃip<sup>h</sup>+/ədiinni/ => [tʃip<sup>o</sup>.ə.di.in.ni] or [tʃi.bə.di.in.ni]  
   'Where is the straw?'

(1a) shows coda neutralization before pause and before another obstruent within a Prosodic Word. (1b) shows that the different types of underlying coda obstruents are realized as an onset when they come before a vowel within the Prosodic Word. (1c) shows that the underlying coda is not realized as an onset across the Prosodic Word boundary. Instead, the coda is neutralized within the Prosodic Word and is realized either as an unreleased coda or as a voiced onset. Korean phonologists have assumed that the neutralized coda is resyllabified as an onset of the following Prosodic Word at the postlexical level. Cho (1987) claims the domain of resyllabification is the Intonational Phrase and Kang (1992) claims it as the Phonological Phrase, larger than the Prosodic Word. However, no phonetic data concerning the domain of resyllabification has been published as far as I know. To determine whether the voiced lenis stop had been resyllabified across the word or across even larger boundaries, the domain of another phonological rule, /l/-flapping was examined as a pilot study.

In Korean, [l] only surfaces in the coda of a syllable, and never as an onset unless the lateral is a geminate. When followed by a vowel, /l/ is resyllabified to an

<sup>1</sup> /p/ becomes [b] intervocalically by the Lenis Stop Voicing rule. The underlying coda /p/ is resyllabified as an onset of the following syllable.

onset of the vowel and appears as a flap [ɾ]. The results show that the resyllabification *can* occur across any word boundaries within an Intonational Phrase (i.e. across the boundary of an Accentual Phrase, the definition of which will be introduced later this section.). Thus, I will assume that any coda lenis stop can be resyllabified to be the onset of the following Accentual Phrase and therefore the following word. I will call this a 'resyllabified' Accentual Phrase initial lenis stop, to distinguish it from the underlying onset lenis stop in Accentual Phrase initial position.

In addition to this asymmetry between syllable onset and coda, however, Korean seems to show an asymmetry at a higher level of prosodic unit. That is, there seems to be a difference between word initial and word final lenis stop in terms of voicing. The word initial voiceless lenis stop in Korean becomes voiced in the middle of the Accentual Phrase but remains voiceless at the Accentual Phrase initial position (Jun 1990a, b). However, in casual speech, informal observation shows that the word final lenis stop becomes voiced most of the time across Accentual Phrase boundaries as well as within the Accentual Phrase. Since lenis stop voicing has been claimed to be a domain span rule in Selkirk's (1986) sense, applying anywhere 'within' the Accentual Phrase, the lenis stop at the end of the Accentual Phrase should be voiceless.<sup>2</sup>

In this paper, I will focus on the voicing of the coda lenis stop at the end of the Accentual Phrase (= at the beginning of the Accentual Phrase after resyllabification), as in {kimpap} {əɭəssni} 'Was the sushi frozen?' (/kimpap/ 'sushi', /əɭ-əss-ni/ 'to freeze-past-Q'). Second, based on the durational relationship between the lenis stop and the adjacent segments in different prosodic positions, I will discuss whether or not Lenis Stop Voicing is a categorical rule. Finally, I will interpret the results in terms of gestural overlapping and reduction based on Browman and Goldstein's (1990) model.

Before introducing the experimental methods, I will briefly introduce the definition of the Accentual Phrase and its relation to the Lenis Stop Voicing rule. The Accentual Phrase is a grouping of Prosodic Words defined on the basis of the tonal pattern of an utterance. In the Seoul dialect, the tonal pattern of the Accentual Phrase is L(H)LH, with the first high optionally appearing when the phrase is longer than four or five syllables. Thus, the salient characteristic of the Accentual Phrase in Seoul Korean is a final rise in pitch. (But, when a phrase is produced with contrastive focus, an initial rise pattern can be found in Seoul, even in a short Accentual Phrase.) In the Chonnam dialect, the characteristic pattern is an initial rise-fall or simple fall, i.e. either LHL or HHL. The choice of pattern is predictable from the laryngeal features of the first segment of the Accentual Phrase: when the segment has either [+spread glottis] (i.e. aspirated consonants and /s/) or [+constricted glottis] (i.e. tensed consonants), the Accentual Phrase has the HHL pattern and otherwise the LHL pattern.

The Accentual Phrase is the comparable level to the Phonological Phrase assumed in Prosodic Phonology (Selkirk 1984, 1986; Nespor and Vogel 1986; Hayes 1989). However, I call it the Accentual Phrase to highlight that its basis is different from what defines the Phonological Phrase. In addition to the syntactic factors emphasized by the prosodic phonologists, the Accentual Phrase is influenced by nonsyntactic and nonlinguistic factors such as focus, speech rate and weight of a phrase (for more detail, see Jun 1993).

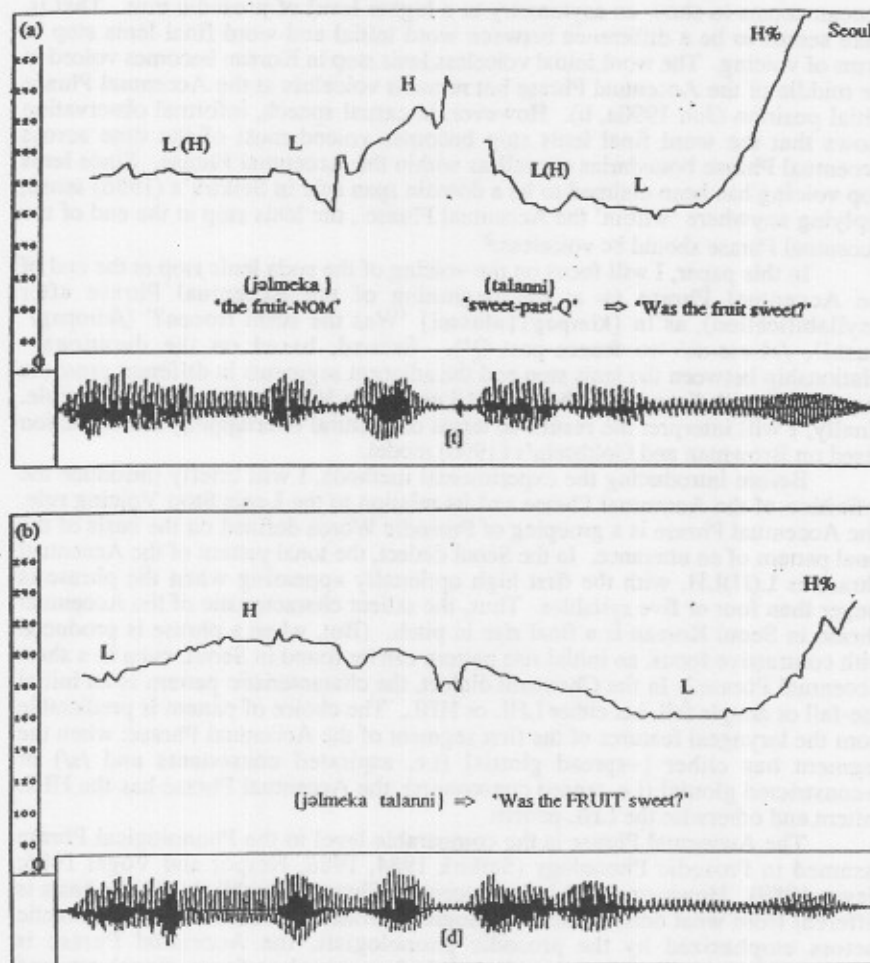
Jun (1990a, b) found that this Accentual Phrase is the domain of Lenis Stop Voicing. That is, the underlyingly voiceless lenis stop is voiced intervocally in

<sup>2</sup> Cho (1987, 1990), Kang (1992) and Silva (1989, 1992) have proposed the Phonological Phrase as the domain of Lenis Stop Voicing rule based on either Selkirk's (1986, 1990) end based theory or Nespor and Vogel's (1986) relation based theory.

the middle of the Accentual Phrase but remains voiceless at the beginning of the Accentual Phrase. Figure 1 illustrates the application of Lenis Stop Voicing in different positions in the Accentual Phrase produced by a Seoul speaker. The X-axis is time dimension and Y-axis is the fundamental frequency,  $f_0$ , value in Hz. (This format will be used for other pitch track figures in this paper.) Figure 1(a)

shows the pitch track and waveform of the sentence (2) produced (a) in two Accentual Phrases and (b) in one Accentual Phrase.

- (2) *jəlme-ka*      *tal-ass-ni*  
 'the fruit - NOM'   'sweet-past-interrogative marker' -> 'Was the fruit sweet?'



**Figure 1** Pitch tracks and waveforms of *jəlme-ka talanni* 'Was the fruit sweet?' in two Accentual Phrasings by Seoul speaker, S2, forming (a) two Accentual Phrases as in *[jəlme-ka] [talanni]* and (b) one Accentual Phrase as in *[jəlme-ka talanni]*.

The Accentual Phrases in Figure 1(a) have a final rise with an initial high being undershot, but the Accentual Phrase in Figure 1(b) has both an initial rise and a final rise. As shown by the absence of the sinusoidal waveform and the broken line on the pitch tracks in Figure 1(a), the Accentual Phrase initial lenis stop is voiceless, [t]. On the other hand, the same lenis stop is voiced, [d], in the middle of the Accentual Phrase as in Figure 1(b). Figure 2 shows the same fact but only differs from Figure 1 in that Figure 2 is produced by a Chonnam speaker, thus having a different verbal ending and an initial rise contour, LHL.

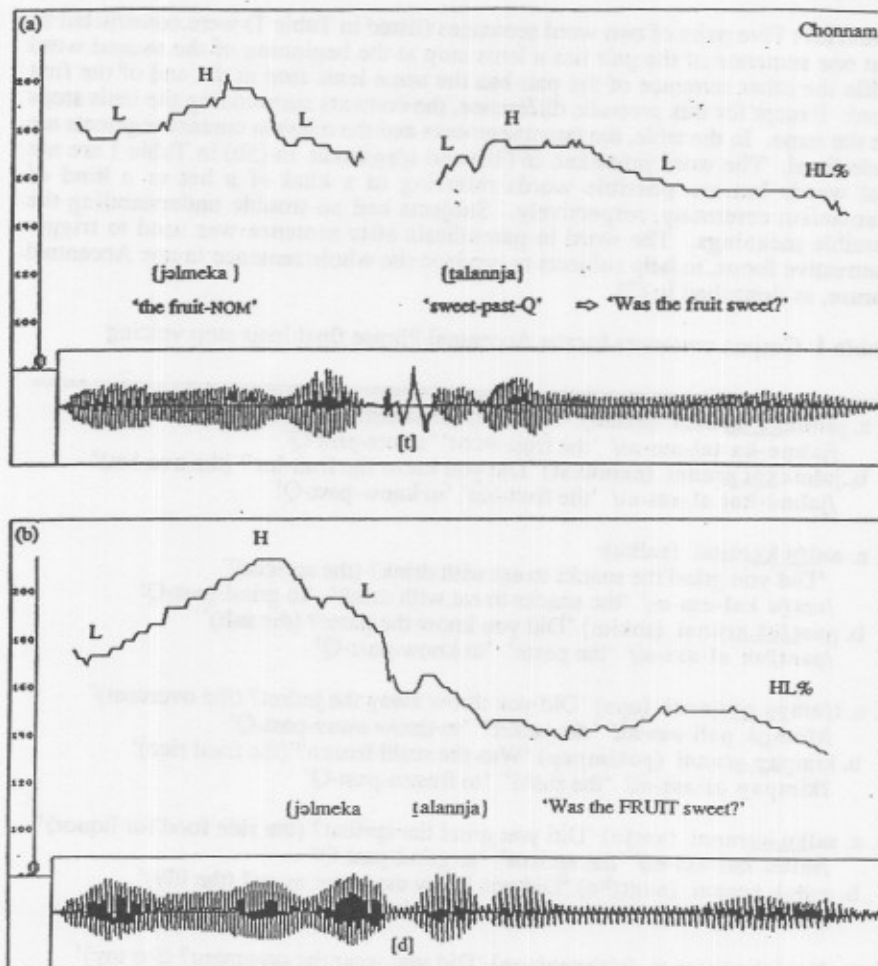


Figure 2. Same as Figure 1, but produced by a Chonnam speaker, C1.

As shown in the figures above, the domain of Lenis Stop Voicing is determined by the tonal pattern of an utterance. That is, a lenis stop at the beginning of an Accentual Phrase remains voiceless, while a lenis stop in the middle of an Accentual

Phrase becomes voiced. To find out the voicing status of the Accentual Phrase final lenis stop, the following experiment was conducted.

## 2. Experimental Methods

**Subjects** : Three Seoul speakers (S1: female, S2: male, and S3: male) and three Chonnam speakers (C1: female, C2: male, and C3: male) were participated in the experiment. All subjects were in their late twenties or early thirties.

**Material** : Five pairs of two word sentences (listed in Table 1) were constructed so that one sentence of the pair has a lenis stop at the beginning of the second word while the other sentence of the pair has the same lenis stop at the end of the first word. Except for this prosodic difference, the contexts surrounding the lenis stops are the same. In the table, the target segments and the relevant context segments are underlined. The word *jəlmekat* in (1b) and *tʃaŋsinkut* in (5b) in Table 1 are not real words but are possible words referring to a kind of a hat or a kind of shamanism ceremony, respectively. Subjects had no trouble understanding the possible meanings. The word in parenthesis after sentence was used to trigger contrastive focus, to help subjects to produce the whole sentence in one Accentual Phrase, as described in (3).

**Table 1** Corpus sentences for the Accentual Phrase final lenis stop voicing

- |   |
|---|
| 1. a. <u>jəlmek</u> a taranni (namu) 'Was the fruit sweet? (the tree)'<br>/jəlmek-a tal-ass-ni/ 'the fruit-NOM' 'sweet-past-Q'  |
| b. jəlmek <u>at</u> aranni (namukat) 'Did you know the fruit-hat? (the tree-hat)'<br>/jəlmek-at al-ass-ni/ 'the fruit-hat' 'to know-past-Q'   |
| 2. a. ant <u>fu</u> karanni (salku)<br>'Did you grind the snacks to eat with drink? (the apricot)'<br>/antfu kal-ass-ni/ 'the snacks to eat with drink' 'to grind-past-Q'                       |
| b. pant <u>fuk</u> aranni (sokim) 'Did you know the paste? (the salt)'<br>/pantfuk al-ass-ni/ 'the paste' 'to know-past-Q'  |
| 3. a. tʃamp <u>a</u> pəɾjənni (opa) 'Did you throw away the jacket? (the overcoat)'<br>/tʃampa pəli-əss-ni/ 'the jacket' 'to throw away-past-Q'   |
| b. kimp <u>ap</u> əɾənni (pokimpap) 'Was the sushi frozen? (the fried rice)'<br>?kimpap əl-əss-ni/ 'the sushi' 'to frozen-past-Q'   |
| 4. a. salk <u>u</u> karanni (antfu) 'Did you grind the apricot? (the side food for liquor)'<br>/salku kal-ass-ni/ 'the apricot' 'to grind-past-Q'   |
| b. suk <u>uk</u> aranni (nantʃo) 'Did you know the water mum? (the lily)'<br>/sukuk al-ass-ni/ 'the water mum' 'to know-past-Q'   |
| 5. a. tʃaŋs <u>inku</u> taranni (tʃaŋnankam) 'Did you wear the ornament? (the toy)'<br>/tʃaŋsinku tal-ass-ni/ 'the ornament' 'to wear-past-Q'   |
| b. tʃaŋs <u>inkut</u> aranni (nerimkut) 'Did you know "tʃaŋsin-shamanism ceremony"?' ('descending shamanism ceremony')<br>/tʃaŋsin-kut al-ass-ni/ 'tʃaŋsin-shamanism ceremony' 'to know-past-Q' |

**Methods** : These sentences were placed in pseudo-random order so that no sentence came after the other sentence from the pair to avoid putting emphasis on



the difference. Seoul and Chonnam dialect speakers were asked to read the whole list in two different Accentual Phrasings 10 times each in normal speech rate. First, they read the whole list of sentences in neutral focus without considering the word in parentheses. In this reading, they nearly always produced the sentence as two Accentual Phrases, one for each word within the sentence. Then they read each sentence a second time putting focus on the first word to contrast it with the word in parentheses by making the whole sentence one Accentual Phrase. To help produce the contrast focus naturally, I asked the subjects to make a new sentence by substituting the contrasting word for the original word. An example is shown in (3). The verbal endings given in Table 1 and other example sentences in this paper were for the Seoul speakers. For speakers of the Chonnam dialect, the dialect form [-nja] was substituted for [-ni].

- (3) Given: /jəlmekataranni/? (namu)  
 'Was the fruit sweet? (the tree)'  
 Read: {jəlmekataranni} {namukataranni}?  
 'Was the fruit sweet or was the tree sweet?'

For each utterance, the target lenis stop and context segments were analyzed for voicing using Kay Sonagraph Model 5500 and the pitch track was checked for the Accentual Phrasing. The durations of the target lenis stop and the following vowel were measured using the spectrogram display. To help measurement, the audio waveform and amplitude were displayed simultaneously in the upper window. In addition, I measured the word medial lenis stop (except for 3(b) in Table 1, where /p/ is produced as [p']) to compare with the duration of the word initial lenis stop. I also measured the word final vowel (underlyingly word final or derived word final after resyllabification), which was the vowel preceding the target lenis stop to see whether the segment shows any difference in duration depending on its position relative to the Accentual Phrase, i.e. Accentual Phrase final or medial. For the target lenis stop and the word medial lenis stop, the duration was measured to include closure duration and any voiceless portion after the release (i.e. VOT). The duration of the vowel preceding the target lenis stop was measured from the point where the first formant of the vowel has a clear amplitude (this mostly matches right after the stop release) to the point where the formant stops (this mostly matches the implosion of the target lenis stop). The duration of the vowel following the target lenis stop was measured from the first formant onset after the stop release to the onset of the flapping.

Next, to examine the domain of flapping for each subject, five sentences containing a word final lateral before a vowel-initial word were given after the list in Table 1. Subjects read each sentence in two Accentual Phrasings ten times each as before and the spectrogram was examined to see whether the word final lateral is produced as a flap. The five sentences are given in Table 2. As in (3) above, the contrasting word is given in the parenthesis.

Table 2 Corpus sentences for flapping

- a. əlluŋmaɭ aranni (tʃoraŋmaɭ) 'Did you know the zebra? (a pony)  
 /əllukmaɭal-ass-ni/ 'the zebra' 'to know-past-Q'
- b. ɔribaɭ aranni (kəwi-paɭ) 'Did you know the duck's foot? (the goose's foot)  
 /oli-paɭal-ass-ni/ 'the duck-foot' 'to know-past-Q'
- c. jaŋmuɭ aranni (kukmuɭ) 'Did you know the medicine water? (the soup)  
 /jakmuɭalassni/ 'the medicine water' 'to know-Q'

d. *kojaŋibaɫ aranni* (kaŋatʃipaɫ) 'Did you know the cat's foot? (the puppy's foot)'  
 /kojaŋi-paɫ al-ass-ni/ 'the cat-foot' 'to know-past-Q'

e. *jəlmeriɫ aranni* (namu) 'Did you know the fruit? (the tree)'  
 /jəlme-lil al-ass-ni/ 'the fruit-ACC' 'to know-past-Q'

### 3. Results and Discussion

#### 3.1. Resyllabification of Lateral

Figure 3 shows spectrograms of example sentence (a) in Table 2 above produced in two Accentual Phrasings: (a) {əɫɫuŋmaɫ aranna} and (b) {əɫɫuŋmaɫ} {aranna} uttered by C2, and (c) {əɫɫuŋmaɫ} {aranni} by S1. The resyllabified flap is shown in (a) and (b) and a lateral is shown in (c) and these are marked by an arrow underneath the spectrogram.

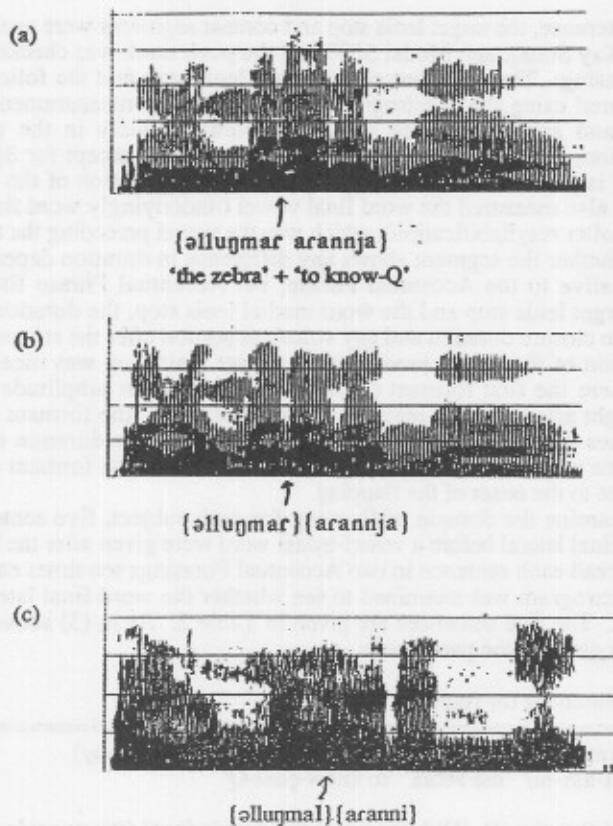


Figure 3. Example spectrograms showing (a) when an Accentual Phrase medial /l/ becomes a flap, (b) when an Accentual Phrase final /l/ becomes a flap, and (c) when an Accentual Phrase final /l/ doesn't become a flap.

As seen in Figure (3a) and (3b), the word final lateral can be resyllabified to be an onset for the following word, showing a flap in both Accentual Phrasings. Thus, we can assume that the word final and phrase final lateral can be resyllabified as the onset of the following word across the Accentual Phrase boundary. The resyllabification across Accentual Phrases occurs in casual speech. In careful and deliberate speech, it does not occur even within an Accentual Phrase. Table 3 shows the percentage of flapping within an Accentual Phrase and across Accentual Phrases for each subject. The percentage is based on 50 tokens.

**Table 3.** Percentage of flapping within and across Accentual Phrases for each subject

Subject	Accentual Phrase medial	Accentual Phrase initial
C1	89.0 %	70.9 %
C2	100 %	98.0 %
C3	81.4 %	80.0 %
S1	66.0 %	63.6 %
S2	82.0 %	68.0 %
S3	83.6 %	73.5 %

For subjects C1, S2, and S3, the word final lateral is flapped more often within the Accentual Phrase than across Accentual Phrases and, for subjects C2, C3, and S1, there seems to be no difference in this regard. Each subject seems to be consistent in their casualness or carefulness in producing a lateral; Subject C2 has flapping most often and S1 least often and this order is consistent within each prosodic position. However, the lateral is not always resyllabified even within the Accentual Phrase. These data suggest that resyllabification is not related very closely to the Accentual Phrase position. But it is clear that the resyllabification *can* occur across Accentual Phrase boundaries.

Generalizing from these utterances, I will assume that any coda consonant type, and specifically the lenis stop, can be resyllabified to be the onset of the following word and therefore the following Accentual Phrase. I will call this a 'resyllabified' Accentual Phrase initial lenis stop, to distinguish it from the underlying onset lenis stop in Accentual Phrase initial position.

### 3.2. Voicing of the Word initial and final lenis stop

Depending on the position of the target lenis stop relative to a Word or an Accentual Phrase, I defined four prosodic positions: onset/A-initial position when the lenis stop is at the beginning of a Word and at the beginning of an Accentual Phrase, onset/A-medial when the lenis stop is at the beginning of a Word but in the middle of an Accentual Phrase, coda/A-initial position when the lenis stop is at the end of a Word but at the beginning of an Accentual Phrase after resyllabification, and coda/A-medial position when the lenis stop is at the end of a Word and in the middle of an Accentual Phrase. The pitch contours and the corresponding segmental realization in onset/A-initial position are what is shown in Figure 1(a) and those in onset/A-medial position are shown in Figure 1(b). Different phrasings and the corresponding segmental realizations in coda/A-initial and coda/A-medial position are shown in Figure 4. The Accentual Phrasing and the voicing of lenis stop in each prosodic position are outlined in (4). The arrow in (4c) indicates a resyllabification. (4c) and (4d) correspond to Figure 4(a) and (b), respectively.

- (4) i. jəlmeka talanni? 'Was the fruit sweet?'  
 a. {jəlmeka} {talanni} => [jəlmega taranni] : onset/A-initial  
 b. {jəlmeka talanni} => [jəlmegadaranni] : onset/A-medial

ii. jəlmekat alanni! Did you know the fruit hat!

- c. {jəlmekat} {alanni} => [jəlmegadaranni] : coda/A-initial  
 d. {jəlmekat alanni} => [jəlmegadaranni] : coda/A-medial

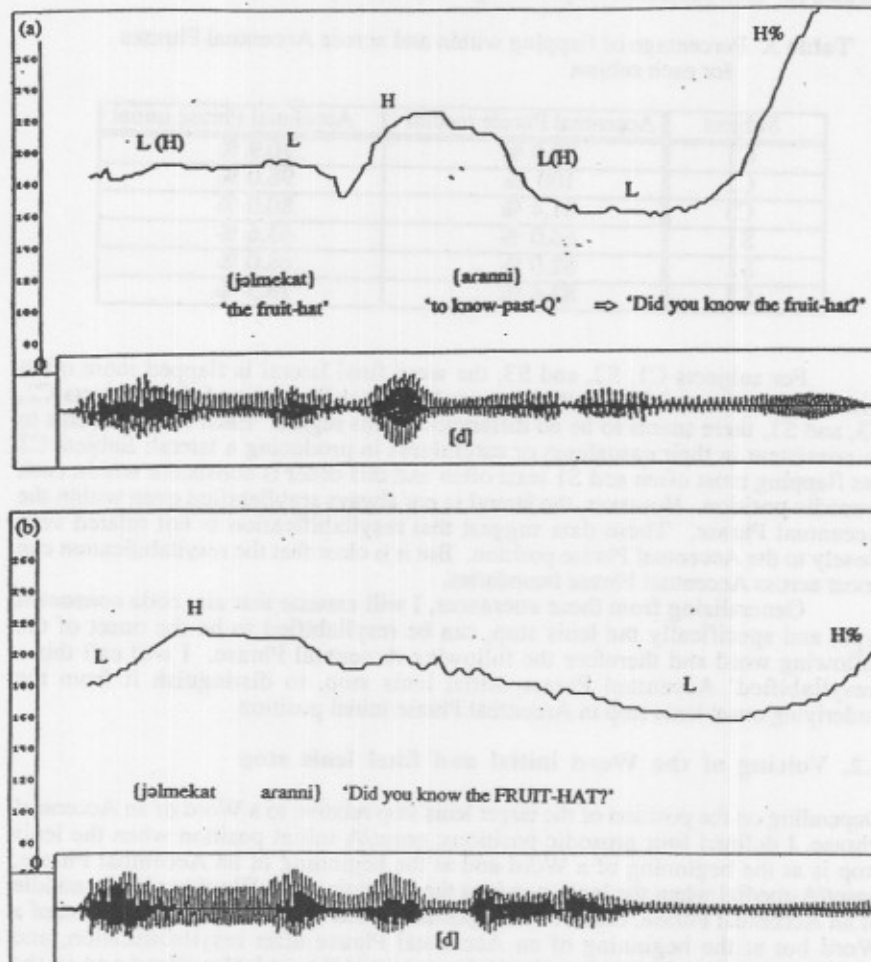


Figure 4. Pitch tracks and waveforms of lenis stop in two prosodic positions: (a) coda/A-initial, (b) coda/A-medial. The sentence is (1b) in Table 1. (speaker: S2)

As expected, the underlying onset or coda /t/ is voiced in Accentual Phrase medial position, (4b) and (4d), whereas the underlying onset /t/ is voiceless at the beginning of the Accentual Phrase, (4a). However, the resyllabified word initial /t/ is still *voiced* as shown in Figure 4(a), i.e. (4c). Thus, even though the tonal pattern of Figure 4(a) is different from those of Figure 1(b) and Figure 4(b), all three are alike in terms of segmental realization. Most of the time, it was hard to distinguish between the type (4b) and (4d) when I was listening without looking at the text. But the type (4c) was easily distinguished from (4b) and (4d) due to the different tonal pattern.

The result of the experiment shows that, as found before, for six subjects, onset stops are mostly voiceless at the Accentual Phrase initial position and voiced in the Accentual Phrase medial position. But word final coda stops are mostly voiced all the time. Out of 300 tokens (5 sentences \* 6 subjects \* 10 repetitions) for each prosodic condition, in general, 5 to 10 % of tokens show an exception to this voicing pattern. (10.67 % voiced at Onset/A-initial position, 4.78% voiceless at Onset/A-medial position, 8.36% voiceless at Coda/A-initial position, and, 4.76% voiceless at Coda/A-medial position.) Figure 5 shows the percentage of voiced versus voiceless lenis stop in four prosodic positions.

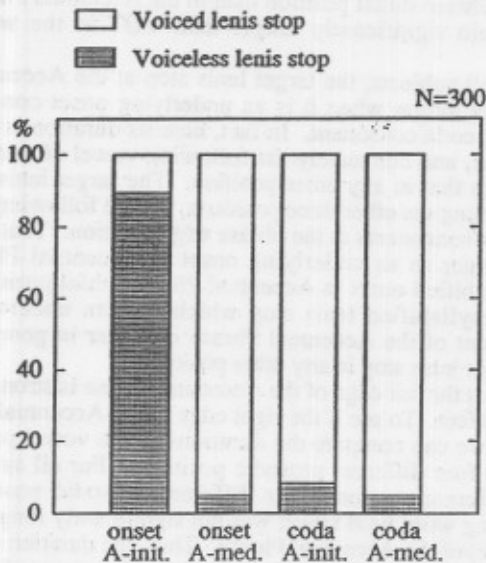



Figure 5. Percentage of voiced versus voiceless lenis stop in four prosodic positions combining data from 6 subjects (N=300).

In summary, though it is not perfect, we can predict most of the voicing data (90 to 95% of occurrences) in terms of the underlying and surface prosodic context of the lenis stop. That is, underlying onset stops are voiceless at the beginning of the Accentual Phrase and voiced in the middle of the Accentual Phrase, whereas underlying coda stops are nearly always voiced.

### 3.3. Duration of lenis stop and adjacent segments in different prosodic positions



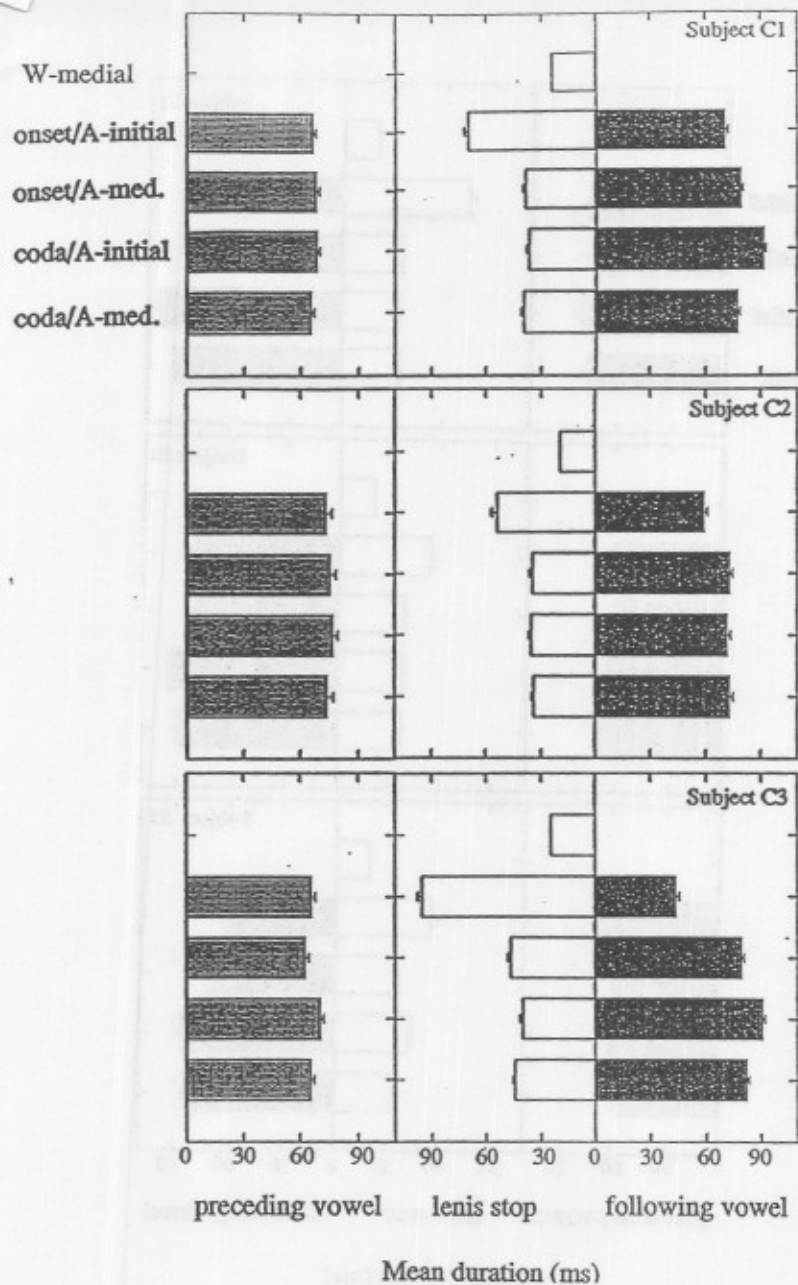
The mean durations of the vowel preceding the target lenis stop, the target lenis stop itself, and the following vowel are plotted in Figure 6 in the four different prosodic positions: onset/A-initial, onset/A-medial, coda/A-initial, and coda/A-medial position. The mean duration of the word medial lenis stop, the lenis stop preceding the target lenis stop, is shown in the first row in the target lenis stop column in Figure 6. Here, the mean value of the word medial lenis stop is only based on the word medial lenis stop between vowels. The error bars indicate the standard error.

For all subjects, there is an effect of Prosodic Word boundary on the duration of lenis stops; the target onset and coda consonants, all of which are at the edges of the word, are substantially longer than the word medial stop. There was also an effect of prosodic phrase boundary; the target onset consonant is substantially longer in Accentual Phrase initial position than in Accentual Phrase medial position. This conforms to the previous results found in Jun (1990a) about the duration of VOT; that is, VOT of word initial aspirated stop was significantly longer in the Accentual Phrase initial position than in the Accentual Phrase medial position which was again significantly longer than VOT in the word medial position.

In addition, for all subjects, the target lenis stop at the Accentual Phrase boundary is substantially longer when it is an underlying onset consonant than when it is an underlying coda consonant. In fact, here its duration is longer than that in any other position, and conversely the following vowel of this position is substantially shorter than that in any other position. The target lenis stop is not significantly different among the other three positions, but the following vowel is in general longer after coda consonants in the phrase edge position. That is, the lenis stop is substantially longer as an underlying onset in Accentual Phrase initial position than as a resyllabified onset in Accentual Phrase initial position, and the vowel following the resyllabified lenis stop which we can understand as the underlying initial segment of the Accentual Phrase is longer in general than the vowel following the target lenis stop in any other position.

Thus, it seems that the left edge of the Accentual Phrase is strong in Korean; it shows a lengthening effect. To see if the right edge of the Accentual Phrase also shows the same effect, we can compare the durations of the vowel preceding the target lenis stop for the four different prosodic positions. For all subjects, there was no significant difference among four different prosodic positions. The duration of the underlying word final vowel was not significantly longer when it is at the end (the right edge) of the Accentual Phrase. Thus, the duration data of word final vowels indicate that it is not necessarily both edges of the prosodic unit which show a segmental lengthening. That is, the boundary effect is not necessarily symmetrical.

Moreover, there are differences among different levels: A segment is very much lengthened at the right boundary of Intonational Phrase (Jun 1992). Therefore, the prosodic boundary effect on the segment is not uniform: Words and Accentual Phrases show a left edge lengthening while Intonational Phrase shows a right edge lengthening. Also the domains and patterns of these lengthening effects are not universally the same: unlike Korean, English has a right boundary effect at both the Word level and Intonational Phrase level (Beckman and Edwards 1990, Crystal and House 1990).



**Figure 6.** The mean duration of the word medial lenis stop, the vowel preceding the target lenis stop, the target lenis stop, and the following vowel in four different prosodic conditions (onset/A-initial, onset/A-medial, coda/A-initial, and coda/A-medial) for each subject.

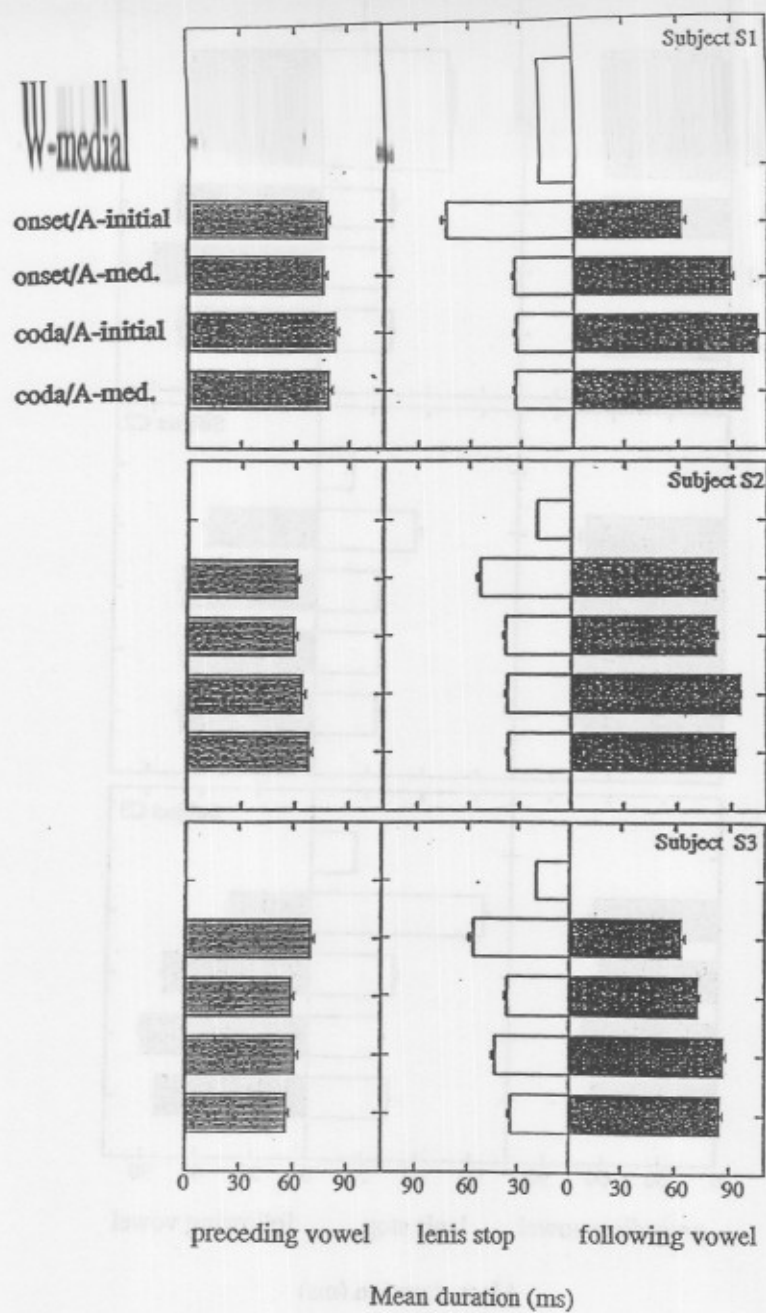


Figure 6. (Continued)



### 3.4. The representation of lenis stop voicing

I have shown in earlier studies that the lenis stop is almost always voiceless at the beginning of the Accentual Phrase and voiced in the middle of the Accentual Phrase. Thus, the Lenis Stop Voicing rule was represented as a domain span rule, which is limited to the Accentual Phrase,  $\alpha$ , as shown in (5).

(5)

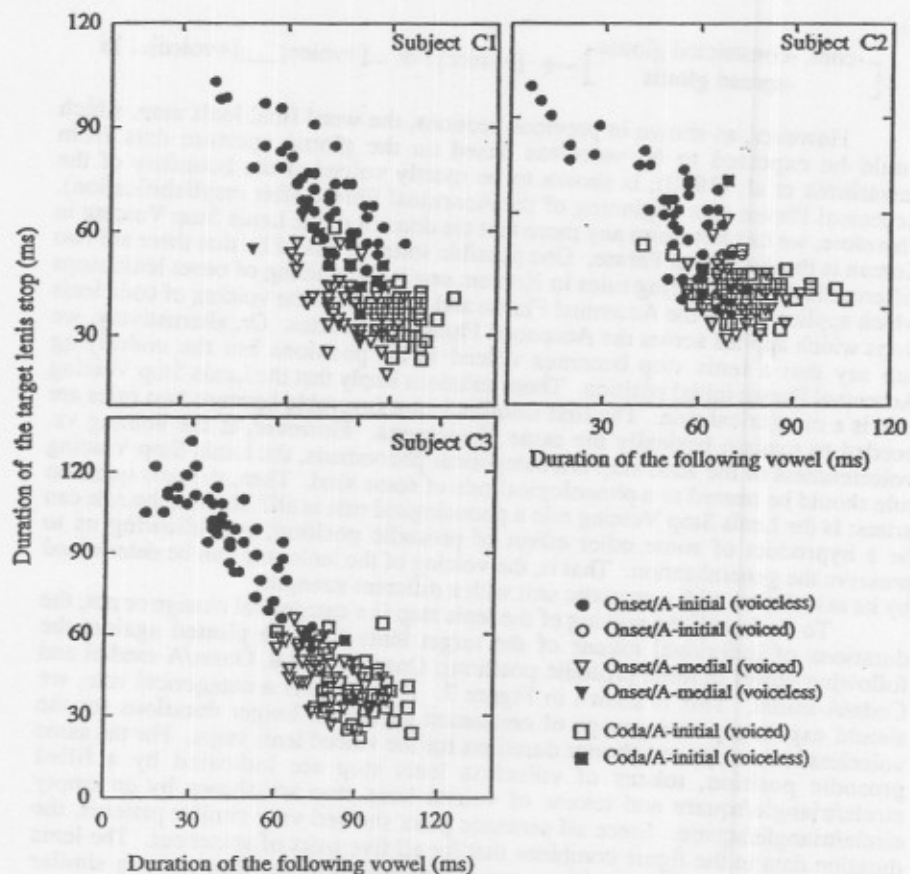
$$\left[ \begin{array}{l} \text{-cont, -constricted glottis} \\ \text{-spread glottis} \end{array} \right] \rightarrow \text{[+voice]} / \alpha \left( \dots \text{[+voice]} \_\_\_\_\_\_ \text{[+voice]} \dots \right) \alpha$$

However, as shown in previous sections, the word final lenis stop, which would be expected to be voiceless based on the glottal aperture data from Sawashima et al. (1980), is shown to be mostly voiced at the boundary of the Accentual Phrase (the beginning of the Accentual Phrase after resyllabification). Therefore, we can not claim any more that the domain of the Lenis Stop Voicing in Korean is the Accentual Phrase. One possible solution would be that there are two different lenis stop voicing rules in Korean: one is the voicing of onset lenis stops which applies within the Accentual Phrase and the other is the voicing of coda lenis stops which applies across the Accentual Phrase boundaries. Or, alternatively, we can say that a lenis stop becomes voiced in all positions but the underlying Accentual Phrase initial position. These solutions imply that the Lenis Stop Voicing rule is a categorical rule. The first solution is not favorable because two rules are needed to explain basically the same phenomena. However, if the voicing vs. voicelessness of the lenis stop is a categorical phenomena, the Lenis Stop Voicing rule should be treated as a phonological rule of some kind. Then, the next question arises: Is the Lenis Stop Voicing rule a phonological rule at all? Rather, the rule can be a byproduct of some other effect of prosodic position, still allowing us to preserve the generalization. That is, the voicing of the lenis stop can be determined by its association with a prosodic unit with a different strength.

To find out if the voicing of the lenis stop is a categorical change or not, the durations of individual tokens of the target lenis stop are plotted against the following vowel in three prosodic positions: Onset/A-initial, Onset/A-medial and Coda/A-initial. This is shown in Figure 7. If the rule is a categorical rule, we should expect separate groups of consonant durations: longer durations for the voiceless lenis stops and shorter durations for the voiced lenis stops. For the same prosodic position, tokens of voiceless lenis stop are indicated by a filled circle/triangle/square and tokens of voiced lenis stop are shown by an empty circle/triangle/square. Since all sentence pairs showed very similar patterns, the duration data in the figure combines that for all five pairs of sentences. The lenis stops in Coda/A-medial positions are not plotted because they show a similar pattern to that of Onset/A-medial position.

For all subjects, there is no clear separation between voiced and voiceless lenis stop duration. Rather, the duration of the lenis stop is negatively related to that of the following vowel: the longer the stop, the shorter the following vowel. That is, it seems that the duration of the lenis stop is trading off with that of the following vowel. Furthermore, no subject shows a clear separation between groups of the data for the different prosodic positions. Although Subject C3 seems to have a better separation between tokens in Onset/A-initial position and the tokens of the other two groups, if we compare voiced tokens with voiceless tokens in the same prosodic position, we can see clearly that the voicing of the lenis stop is predicted by the relative duration of the lenis stop and the following vowel: i.e.

longer stops followed by shorter vowels tend to be voiceless and shorter stops followed by longer vowels tend to be voiced.



**Figure 7.** The duration of the target lenis stop plotted against the following vowel in three different prosodic positions for each subject: onset/A-initial, onset/A-medial and coda/A-initial position. Tokens of voiced lenis stop are indicated by a filled circle/triangle/square, and tokens of voiceless lenis stop are indicated by an empty circle/triangle/square.

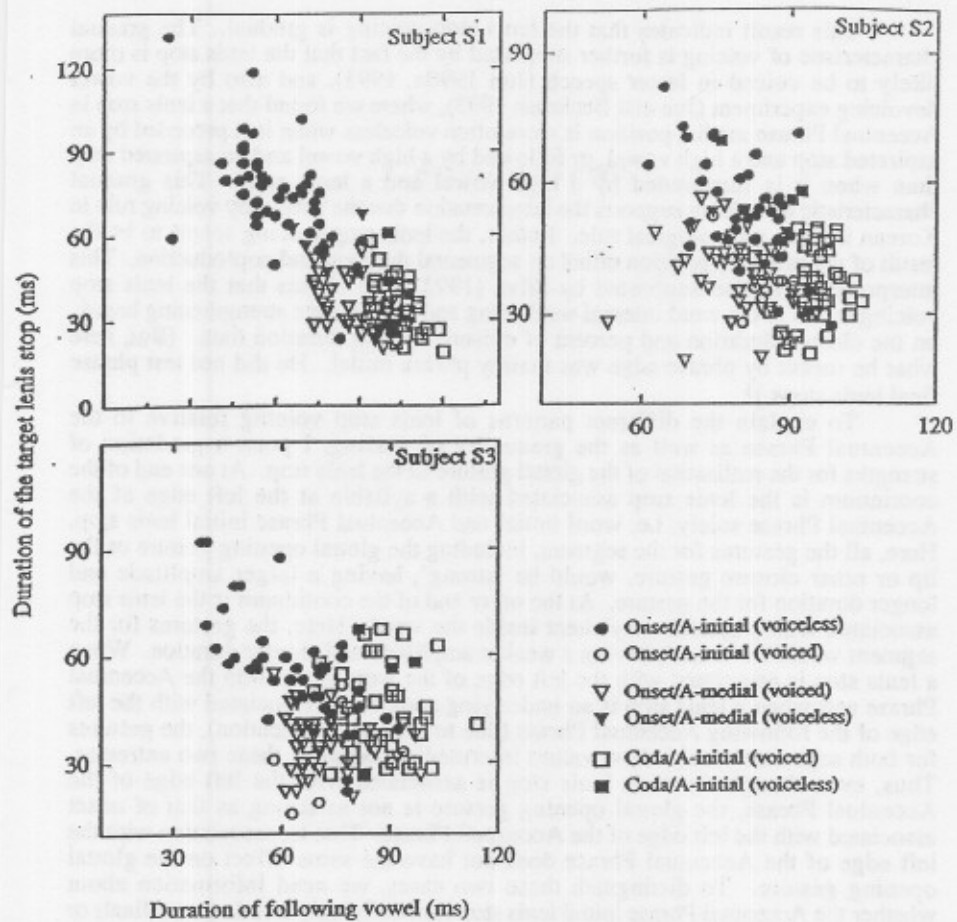



Figure 7. (Continued)

This result indicates that the lenis stop voicing is gradual. The gradual characteristic of voicing is further supported by the fact that the lenis stop is more likely to be voiced in faster speech (Jun 1990a, 1993), and also by the vowel

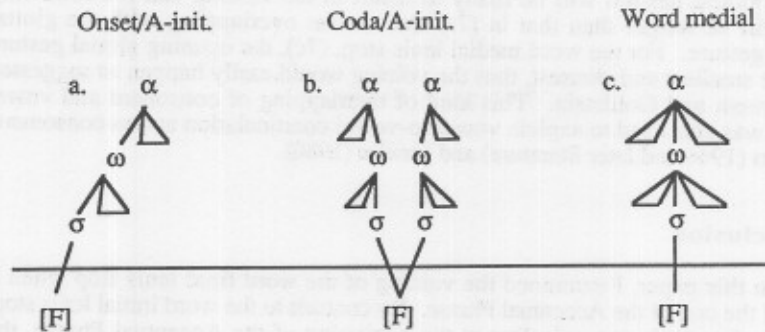


devoicing experiment (Jun and Beckman 1993), where we found that a lenis stop in Accentual Phrase medial position is more often voiceless when it is preceded by an aspirated stop and a high vowel, or followed by a high vowel and an aspirated stop than when it is surrounded by a high vowel and a lenis stop. This gradual characteristic of voicing supports the interpretation that the lenis stop voicing rule in Korean is not a phonological rule. Rather, the lenis stop voicing seems to be the result of the prosodic position effect on segmental duration and coproduction. This interpretation is also supported by Silva (1992) who claims that the lenis stop voicing is due to the word internal weakening and phrase edge strengthening based on the closure duration and percent of closure voicing duration data. (But, here what he meant by phrase edge was mainly phrase initial. He did not test phrase final lenis stops.)<sup>3</sup>

To explain the different patterns of lenis stop voicing relative to the Accentual Phrase as well as the graduality of voicing, I posit a gradation of strengths for the realization of the glottal gesture of the lenis stop. At one end of the continuum is the lenis stop associated with a syllable at the left edge of the Accentual Phrase solely, i.e. word initial and Accentual Phrase initial lenis stop. Here, all the gestures for the segment, including the glottal opening gesture or the lip or other closure gesture, would be 'strong', having a larger amplitude and longer duration for the gesture. At the other end of the continuum is the lenis stop associated with a syllable anywhere inside the word. Here, the gestures for the segment would be 'weak', having a weaker amplitude and shorter duration. When a lenis stop is associated with the left edge of the word but within the Accentual Phrase and when a lenis stop is an underlying coda but is associated with the left edge of the following Accentual Phrase (due to the resyllabification), the gestures for both segments would show values intermediate between these two extremes. Thus, even though the coda lenis stop is associated with the left edge of the Accentual Phrase, the glottal opening gesture is not as strong as that of onset associated with the left edge of the Accentual Phrase. That is, association with the left edge of the Accentual Phrase does not have the same effect on the glottal opening gesture. To distinguish these two cases, we need information about whether the Accentual Phrase initial lenis stop is underlyingly a coda (word final) or an onset (word initial). Schematic representations of the prosodic structures conditioning the two extremes of the continuum and the coda/A-initial type lenis stop are shown in (6). Here,  $\alpha$  is an Accentual Phrase,  $\omega$  a Prosodic Word, and [F] is the bundle of features specifying the lenis stop. The horizontal line separates the prosodic specification plane from the associated segmental features. (6a) is the representation for the Accentual Phrase initial onset lenis stop. (6b) is the representation for the coda stop resyllabified across the Accentual Phrase boundary. (6c) is the representation for the word medial lenis stop. To represent the different voicing pattern of the underlying onset Accentual Phrase initial and the underlying coda Accentual Phrase initial lenis stop, [F] is associated with one  $\alpha$  in (6a) but two  $\alpha$ s in (6b).

<sup>3</sup> His PE (phrase initial) category is determined based on syntactic structure of a sentence. Thus his PE is not necessarily the same as my Accentual Phrase initial. Therefore, since some of his WE (word initial) or PE could be my Accentual Phrase initial or Accentual Phrase medial, I can't compare his results with mine in terms of voicing related duration data.

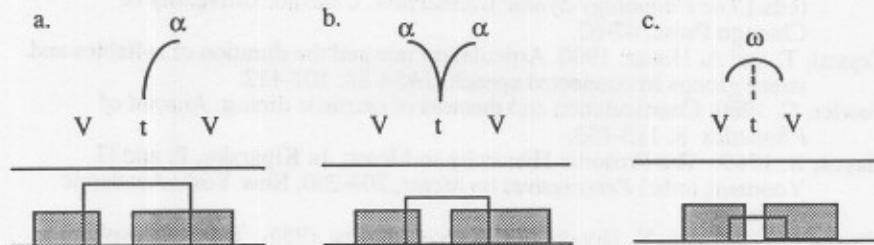
(6) Schematic representations of the prosodic structure conditioning the two extremes of the continuum and the coda/A-initial type lenis stop



Browman and Goldstein (1990) explain the intervocalic voicing assimilation as a reduction in the magnitude of the glottal opening-and-closing gesture responsible for the voicelessness. That is, if the magnitude of the opening is reduced sufficiently, devoicing might not take place at all. Based on data from Japanese (Hirose et al. 1985), where the separation between the vocal folds at the point where voicing ceases at the beginning of an intervocalic voiceless stop is much larger than at the point where voicing begins again at the end of the stop, they suggest that if the magnitude of the abduction gestures were slightly reduced, the critical value of vocal fold separation for devoicing might never be reached.

However, in addition to the different amplitude of the glottal gesture, the negative gradual relationship between lenis stop and the following vowel shown in Figure 7 suggests that there is a gestural overlapping between the lenis stop's glottal opening gesture and the following vowel's glottal closing gesture. That is, the different degrees of overlapping between the glottal opening or closing gestures and the different degrees of amplitude of the glottal gesture would produce the gradual voicing output. The hypothetical gestural score for a lenis stop, here /t/, in different prosodic positions is given in (7). Only the glottal tier is shown. The height of the box indicates degree of opening (aperture) or closing (closure) of the glottal gesture and the width of the box indicates the gesture's duration. The white box is for the glottal opening gesture and the shaded boxes are for the glottal closing gestures.

(7) Hypothetical score of overlapping glottal gestures



For the Accentual Phrase initial lenis stop, (7a), the opening gesture would be larger and longer, overlapping and hiding the vowel's glottal closing gesture,

while for the resyllabified phrase initial lenis stop, (7b), the opening gesture would be smaller and shorter and overlapping less with the following vowel. This weaker opening glottal gesture will be likely to result in the voicing and the following vowel will be longer than that in (7a) due to less overlapping with the glottal opening gesture. For the word medial lenis stop, (7c), the opening glottal gesture would be smallest and shortest, thus the voicing would easily happen as suggested by Browman and Goldstein. This kind of overlapping of consonant and vowel gestures was also used to explain vowel-to-vowel coarticulation across consonants in Ohman (1966 and later literature) and Fowler (1980).

#### 4. Conclusion

In this paper, I examined the voicing of the word final lenis stop when it comes at the end of the Accentual Phrase. By contrast to the word initial lenis stop, which is almost always voiceless at the beginning of the Accentual Phrase, the word final lenis stop was voiced at the resyllabified Accentual Phrase initial position. The data show that the voicing of lenis stop depends on its duration relative to the following vowel and this duration is determined by their position relative to the prosodic contexts. Therefore, I proposed that the Lenis Stop Voicing rule in Korean is not a phonological rule, but is a byproduct of some other effect of prosodic position on the gestural amplitude and overlapping, thus producing a continuum of voicing. To distinguish the different duration pattern of the lenis stop, thus the different voicing pattern of the lenis stop, I suggested different prosodic representations utilizing the coda/onset information.

#### References

- Beckman, M. and J. Edwards. 1990. Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston and M. Beckman (eds.) *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge, England: Cambridge University Press, 152-178.
- Browman, C. and L. Goldstein. 1990. Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. Beckman (eds.) *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge, England: Cambridge University Press, 341-376.
- Cho, Y. Y. 1987. The Domain of Korean Sandhi Rules. Paper presented at the 62nd LSA meeting.
- Cho, Y. Y. 1990. Syntax and Phrasing in Korean. In S. Inkelas and D. Zec (eds.) *The Phonology-Syntax Connection*. Chicago: University of Chicago Press, 47-62.
- Crystal, T. and A. House. 1990. Articulation rate and the duration of syllables and stress groups in connected speech. *JASA* 88: 101-112.
- Fowler, C. 1980. Coarticulation and theories of extrinsic timing. *Journal of Phonetics* 8: 113-133.
- Hayes, B. 1989. The Prosodic Hierarchy in Meter. In Kiparsky, P. and G. Youmans (eds.) *Perspectives on Meter*, 203-260, New York: Academic Press.
- Hirose, H., S. Niimi, K. Honda, and M. Sawashima. 1985. The relationship between glottal opening and transglottal pressure difference during consonant production. *Annual Bulletin of RILP* 19: 55-64.
- Jun, S.-A. 1990a. The Domains of Laryngeal Feature Lenition Effects in Chonnam Korean. Presented at the 119th meeting of the ASA, Baltimore.

- Jun, S.-A. 1990b. The Prosodic Structure of Korean - in terms of voicing. In E-J. Baek (ed.) *Proceedings of the Seventh International Conference on Korean Linguistics*, Vol. 7. Univ. of Toronto Press.
- Jun, S.-A. 1992. The Domain of Nasalization and the Prosodic Structure in Korean. In H. Sohn (ed.) *Korean Linguistics 7*: 11-29.
- Jun, S.-A. 1993. *The Phonetics and Phonology of Korean Prosody*. Ph.D. dissertation. The Ohio State University.
- Jun, S.-A. and M. Beckman. 1993. A gestural-overlap analysis of vowel devoicing in Japanese and Korean. Paper presented at the 67th LSA Meeting, Los Angeles, CA.
- Kagaya, R. 1974. A Fiberscopic and Acoustic Study of the Korean Stops, Affricatives and Fricatives. *Journal of Phonetics 2*: 161-180.
- Kang, O. 1992. *Korean Prosodic Phonology*. Ph.D. dissertation. University of Washington.
- Keating, P., W. Linker, and M. Huffman. 1983. Patterns in allophone distribution for voiced and voiceless stops. *Journal of Phonetics 11*: 277-290.
- Nespor, M. and I. Vogel. 1986. *Prosodic Phonology*. Dordrecht: Foris.
- Ohman, S. 1966. Coarticulation in VCV utterances: spectrographic measurements. *JASA 41*: 310-320.
- Oller, D. K. 1973. The effect of position in utterance on speech segment duration in English. *JASA 54*: 1235-1247.
- Pierrehumbert, J. and D. Talkin. 1992. Lenition of /h/ and Glottal Stop. In G. Docherty & D. R. Ladd (eds.) *Papers in Laboratory Phonology II: Gestures, Segment, Prosody*, Cambridge, England: Cambridge University Press, 90-116.
- Sawashima, M., H-S. Park, K. Honda, and H. Hirose. 1980. Fiberscopic study on laryngeal adjustments for syllable-final applosives in Korean. *Ann. Bull. RILP*, No. 14: 125-138.
- Selkirk, E. 1984. *Phonology and Syntax: The Relation between Sound and Structure*, Cambridge, MA: MIT Press.
- Selkirk, E. 1986. On Derived Domains in Sentence Phonology. *Phonology Yearbook 3*: 371-405.
- Silva, J. D. 1989. Determining the Domain for Intervocalic Stop Voicing in Korean. In S. Kuno et al. (eds.) *Harvard Studies in Korean Linguistics III.*, Cambridge, MA: Harvard Univ. Press.
- Silva, J. D. 1992. *The Phonetics and Phonology of Stop Lenition in Korean*. Ph.D. dissertation. Cornell University.

## Is there 'dephrasing' of the accentual phrase in Japanese?\*

Kikuo Maekawa

kikuo@tansei.cc.u-tokyo.ac.jp

**Abstract:** An experiment was carried out in order to examine two putative cases of 'dephrasing' of the accentual phrase in Japanese. The result revealed that it was possible to detect the accentedness of seemingly 'dephrased' accentual phrases in most of the cases. Although  $f_0$  contours of the accentual phrases in question are quasi-linear, we can detect their accentedness through the statistical examination of slope and intercept values of the regression lines fitted to the  $f_0$  contours.

### 1. Introduction

#### 1.1. *Accentual phrases and their dephrasing*

In this paper, I will examine the notion of 'dephrasing' of the accentual phrase in Standard (Tokyo) Japanese. Experimental evidence will indicate that, contrary to current standard theoretical assumptions, it is possible to detect the accentedness of seemingly 'dephrased' accentual phrases. In this section I will review briefly the current theoretical assumptions about 'dephrasing'.

Standard Japanese has the system of pitch accent. Japanese accent differs considerably from the pitch accent of languages like English (and many other European languages) in that it is specified almost exclusively at the level of the lexicon. The accentual phrase of Japanese is usually defined as the domain of two independent phonological/phonetic events. It is the domain of at most one accent and it is also the domain of the phrase initial rise which demarcates the boundary of two successive accentual phrases. Because accent is paradigmatically contrastive in Japanese, Japanese accentual phrases can be classified into accented and unaccented ones according to their accentedness. In its physical realization, an accented accentual phrase is characterized by a sharp fall of fundamental frequency ( $f_0$ ) at the designated accent location, whereas an unaccented accentual phrase

---

\*This study was carried out during my stay at OSU. I would like to express my gratitude to Mary Beckman who made my stay possible and provided many invaluable comments to this and other studies that I conducted at OSU. My gratitude goes also to Osamu Fujimura and Masashi Sawada who gave me various comments in the course of the study. My stay in the U.S. was supported by the Ministry of Education, Japan.



does not possess such a fall, and is instead characterized by a relatively gradual pitch fall spanning the whole time portion after its initial rise.

An important assumption which virtually all phonological analyses of Japanese intonation hold is that two or more accentual phrases can be dephrased and merged into one thereby deleting all accents except the left-most one (e.g. McCawley, 1968; Poser, 1984). Under this standard assumption, the effect of dephrasing can be shown schematically as in Figure 1. Dephrasing of unaccented plus unaccented accentual phrases results in one long unaccented accentual phrase (Case 1 of Figure 1), whereas if at least one of the component phrases is accented the resulting phrase is a long accented accentual phrase (Cases 2-4). It is important to note that case 4 differs considerably from all the rest because in this case dephrasing implies the tonal deletion of the second accent. This is one of the rare instances in Japanese where a lexical accent is deleted post-lexically, and it is this case that will be examined in this paper. Two more points are to be noted here. First, the standard assumption predicts that the accentual phrase resulting from the dephrasing of accented and unaccented accentual phrases (Case 3) and the one resulting from the dephrasing of two accented accentual phrases (Case 4) are exactly the same. Second, when two accented accentual phrases are concatenated (i.e. not dephrased), the phonetic prominence of second accent (most notably in terms of its peak height) is usually reduced by the influence of the preceding accent. This effect is known as downstep or catathesis. (Poser, 1984; Pierrehumbert & Beckman, 1988).

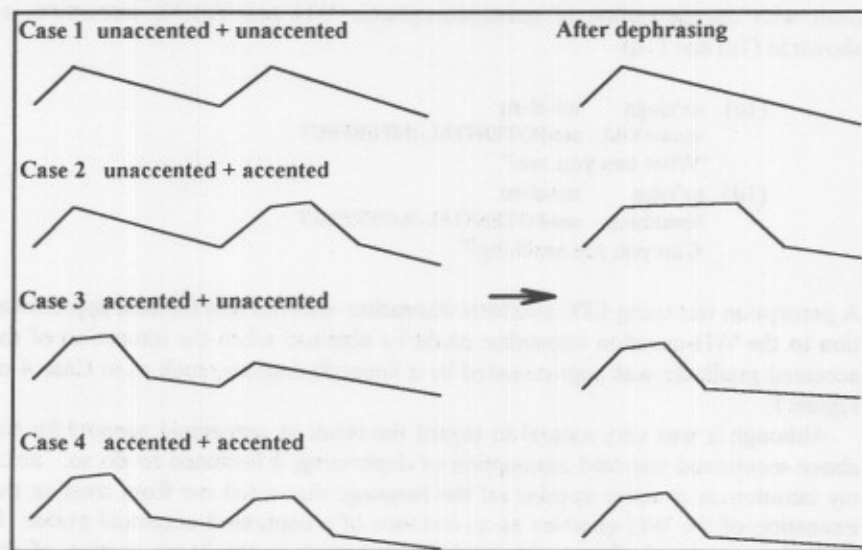


Figure 1. Schematic representation of the effect of dephrasing.

## 1.2. Reported instances of 'dephrasing'

There are two instances of dephrasing mentioned in recent experimental phonological studies of Japanese intonation which will have direct relevance to the present experiment. The first one is dephrasing occurring in post-focus position; when an accented accentual phrase is emphasized due to some pragmatic/discourse factor, the accentual phrase which immediately follows the focused accented accentual phrase seems to be dephrased quite often. Beckman & Pierrehumbert (1986: 265) compared two f0 tracks of (a) *UMA'I mame'wa arimase'n* 'There are no tasty beans' and (b) *UMA'I amewa arimse'n* 'There is no tasty candy' and concluded that "the tonal contours are identical because the grouping has deleted the lexical accent in *mame'*." (Capitals show emphasis and an apostrophe denotes the location of accent.) In this case, the effect of dephrasing can be shown as in (1a) and (1b), where  $[\ ]_{\alpha}$  denotes an accentual phrase.

- (1a) [uma'i] $_{\alpha}$  [mame'-wa] $_{\alpha}$  [ari-mase'-n] $_{\alpha}$   
 tasty beans-TOPIC be-POLITE-NEG  
 => [uma'i mame-wa] $_{\alpha}$  [arimase'n] $_{\alpha}$
- (1b) [uma'i] $_{\alpha}$  [ame-wa] $_{\alpha}$  [ari-mase'-n] $_{\alpha}$   
 tasty candy-TOPIC be-POLITE-NEG  
 => [uma'i ame-wa] $_{\alpha}$  [arimase'n] $_{\alpha}$

In this respect, my own study reported in Maekawa (1991) is to be noted, since it seems to be the only study that examined the perceptual aspect of this issue. It dealt with the distinction of quasi-homophonic WH and Yes-No intonation, as shown in (1c) and (1d).

- (1c) na'ni-ga mi-e'-ru  
 what-NOM see-POTENTIAL-IMPERFECT  
 'What can you see?'
- (1d) na'nika mi-e'-ru  
 Something see-POTENTIAL-IMPERFECT  
 'Can you see anything?'

A perception test using LPC synthetic intonation showed that the best approximation to the WH-question intonation could be obtained when the intonation of the accented predicate was approximated by a linear f0 contour, such as in Case 4 of Figure 1.

Although it was very natural to regard the result as perceptual support for the above mentioned standard assumption of dephrasing, I hesitated to do so, since my intuition as a native speaker of the language dissuaded me from treating the intonation of the WH question as an instance of a dephrased accentual phrase. It seemed to me perfectly possible to hear an accent in the linear portion of the intonation, even in the synthetic stimuli.<sup>1</sup> The experiment which will be presented

<sup>1</sup> The conclusion of Maekawa (1991) was that the intonational difference between WH and Yes-

in section 2.1 below will address this problem and I will propose a solution to it.

The second instance of dephrasing mentioned in the recent experimental phonological literature is the one represented by Poser's (1984: 142ff) treatment of auxiliary verb *mi'ru*. He treated the intonational difference between the two instances of *yonde miru* ('try reading' or 'read and see') as a difference in accentual phrasing (his 'minor phrase').

(1e) [yo'n-de mi-ru]<sub>α</sub>

read-GER see

'try reading'

(1f) [yo'n-de]<sub>α</sub> [mi'-ru]<sub>α</sub>

read-GER see

'read and see'

According to Poser, *mi'ru* in (1e), which functions as an auxiliary verb, is dephrased by the principle of *minimal minor phrase formation* which states roughly that the prosodic word consisting of a content word and any following function word should form a minor phrase. By this account *mi'ru* in (1f) is not dephrased since this is a full verb. Poser's account is not specific to auxiliary verbs; the same account is used for case markers and other particles as well as conjunctions and the copula (see Poser, 1984: 148).

Poser's account of auxiliary verbs was criticized in Kubozono (1993). Based on the visual inspection of larger amount of f0 data<sup>2</sup>, Kubozono found that there was a considerable amount of variation in the physical realization of the accent in auxiliary verbs as well as auxiliaries like *daro'o* 'perhaps' and two morae long particles like *ma'de* 'until' and *yo'ri* 'rather than'.<sup>3</sup> Kubozono classified the allophonic variants observed in his study into the three types shown in Figure 2.

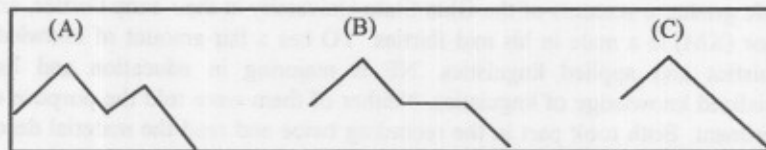


Figure 2. Classification of the realizations of two consecutive accented accentual phrases. Adopted from Kubozono (1993: 104).

Kubozono's type A is the full realization of the auxiliary verb (or other particles) as an independent accentual phrase (his 'minor phrase'), with the second accent (on the auxiliary) downstepped. Type B is what Kubozono calls 'total downstep' following the terminology used by Poser, who adopted the term from Pierrehumbert's (1980) analysis of the English H\*L pitch accent. Type C is the

No questions should be attributed to a difference in intermediate phrasing.

<sup>2</sup> Poser's f0 data was taken from only one speaker, while Kubozono analyzed four.

<sup>3</sup> Auxiliaries like *daro'o* are syntactically different from auxiliary verbs primarily in that they do not inflect, and secondarily because they don't have full-verb counterparts.

expected form of dephrasing as predicted by Poser. Kubozono's conclusion is that apparent 'dephrasing' of auxiliary verbs and other particles is not the case of phonological dephrasing (his 'accent reduction') but it is rather the weakening of an accent caused by the effect of downstep at the phonetic level. It is important to note that Kubozono states at the end of his discussion that it is not unrealistic to suppose that type C "represents an extreme case of downstep, a case where the second accented element is totally reduced by the preceding accented element ---so much so that neither the phrase-initial *F0* rise nor the accent of the second element is now realized as the phonetic output." (p.113, italics mine.)

Kubozono's account is in accordance with native speakers' intuition; in fact the same view was presented by Japanese linguists much earlier back in late 1960s based fully on impressionistic observations. (see Wada 1969 among others, and also Fujimura 1991 for relevant issues.) However, it seems that Kubozono did not explore the full possibility of his proposal. If we follow his view and push it one step further, couldn't we expect that some remaining trace of the accent is to be found even in the type C realization, whose accent Kubozono regarded to be completely unrealized? This point will be examined in section 2.2.

## 2. Experiment

In this section I will show the results of two experiments on the dephrasing of accentual phrases in Japanese. One of them was concerned with dephrasing in post-focus position and the other one was concerned with dephrasing of auxiliary verbs. The experimental material used here is a part of larger data set which awaits complete analysis. At present, two native speakers of Tokyo Japanese plus the author have taken part in the recording, and a few more speakers are expected to be recorded. Both of the speakers —abbreviated as YO and NF below— are female graduate students of the Ohio State University in their early thirties, and the author (KM) is a male in his mid thirties. YO has a fair amount of knowledge in linguistics and applied linguistics. NF is majoring in education and has no specialized knowledge of linguistics. Neither of them were told the purpose of the experiment. Both took part in the recording twice and read the material described below in slightly different ways.

The data set consisted of thirty-four sentences; each one of them was either a question or an answer to a previously uttered question. Each pair of related question and answer was printed on a separate index card and the cards were randomized every time before new recording session started. Every speaker read each card at least ten times. Recording was conducted in a sound booth in a quasi-conversational setting, i.e. a dummy speaker —actually KM, the author— took part in the recording as an interlocutor, and the target speaker was asked to read the pairs of question/answer sentences interchanging his/her role as questioner and respondent with that of dummy speaker. This setting was devised after the first recordings of YO and NF in order to elicit as natural utterances as possible, since in their first recordings sentences prepared for post-focus dephrasing showed considerable variation ranging from Kubozono's type A through C, of which only

type C is relevant for the purpose of the current experiment. Judging from the impressionistic as well as quantitative criterion described below, the new setting was an effective one. All data were taken from the second recordings.

### 2.1. 'Dephrasing' in post-focus position

Two pairs of question/answer sentences were used for this experiment. The WH words in the questions are supposed to be focused in their physical realization. The question mark at the end of sentence stands for the final rise for question rendition.

- (2a) ho'n-o da're-ga yo'n-de-iru'-no?  
book-OBJ who-NOM read-GER-PROG-Q  
'Who is reading the book?'
- (2b) ho'n-wa yu'mi-ga yo'n-de-iru  
book-TOPIC Yumi-NOM read-GER-PROG  
'It is Yumi who is reading the book.'
- (3a) ke'n-o da're-ga yon-de-iru'-no?  
Ken-OBJ who-NOM call-GER-PROG-Q  
Who is calling Ken?
- (3b) ke'n-wa yu'mi-ga yon-de-iru  
Ken-TOPIC Yumi-NOM call-GER-PROG  
'It is Yumi who is calling Ken.'

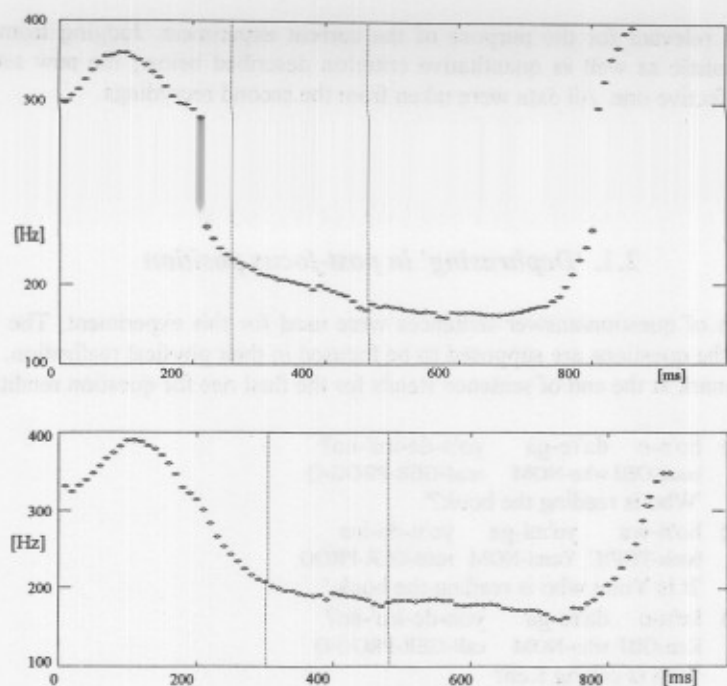
The crucial difference between (2) and (3) is in the accentedness of the predicate verb stems. Under the standard assumption, the accent of the verb in (2a) should be deleted if dephrasing takes place between the second and the third constituents of the sentence due to the intrinsic focus of the WH-word. Hence the resulting intonation contour is expected to be the same as the one observed for (3a), which is under the same effect of dephrasing. The expected prosodic structure shared by (2a) and (3a) under the effect of dephrasing is [da'rega yondeiruno]<sub>α</sub>.<sup>4</sup>

Figure 3 shows typical f<sub>0</sub> contours of the relevant portions of (2a) and (3a) as uttered by speaker YO.<sup>5</sup> Two vertical lines in each panel are set to the beginning and the end of the verb stem defined by spectrographic investigation. The left edge of the stem was defined as the mid point of the second formant transition from /j/ to /o/, and the right edge was defined as the beginning of the vowel /e/ of the gerundial /de/.<sup>6</sup> At first glance, comparison of the two f<sub>0</sub> contours seems to provide good support for the standard assumption of dephrasing since neither a phrase initial rise nor a local f<sub>0</sub> fall due to the accent can be seen in the delimited time portion. For both accented and unaccented verb stems, the whole time portion is associated with a quasi-linear f<sub>0</sub> contour which constitutes a part of simple

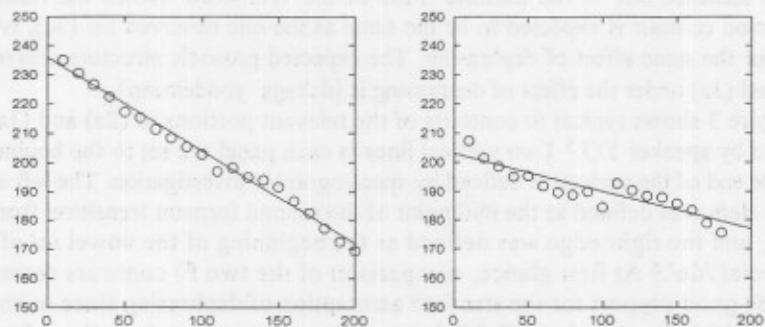
<sup>4</sup> Note it is assumed that the accents at the end of /iru/ in (2) and (3) are both deleted under the effect of dephrasing.

<sup>5</sup> All f<sub>0</sub> tracking was done by the 'formant' program of ESPS with the framelength of 10ms.

<sup>6</sup> I have excluded the segment of /j/ from the acoustic definition of verb stem because in some speakers' f<sub>0</sub> contours the L of bitonal HL accent was realized later in time and found itself in the time segment of /j/.



**Figure 3.** Typical  $f_0$  contours of (2a) and (3a) by YO. Two vertical lines on each panel correspond to the left and right edges of verb stem defined by acoustic segmentation. Part of the  $f_0$  contours preceding WH-words were omitted. The top panel shows (2a) *da'rega yo'ndeiruno?* The bottom panel shows (3a) *da'rega yondeiruno?*.



**Figure 4.** Example of line fitting. Lines were fitted to the  $f_0$  data points extracted from the verb stems of (2a) and (3a) by the least-squares method. Here, the same utterances as in Figure 3 were used for examples. The left panel is the accented verb stem (2a). The right panel is unaccented (3a). For each panel, the abscissa stands for time in ms, and the ordinate stands for  $f_0$  in hertz.

**Table 1: RMS prediction error of (2a) and (3a) for speaker YO**

Sentence	Verb stem	N	Mean	SD	<i>t</i> -test (two-tailed)
(2a)	Accented	172	2.197	1.926	$t = -1.586$ , (D.F.=341.2)
(3a)	Unaccented	178	2.556	2.299	P=0.114

**Table 2: RMS prediction error of (2a) and (3a) for speaker KM**

Sentence	Verb stem	N	Mean	SD	<i>t</i> -test (two-tailed)
(2a)	Accented	191	1.864	1.817	$t = -1.733$ , (D.F.=374.3)
(3a)	Unaccented	193	2.213	2.121	P=0.084

**Table 3: RMS prediction error of (2a) and (3a) for speaker NF**

Sentence	Verb stem	N	Mean	SD	<i>t</i> -test (two-tailed)
(2a)	Accented	186	4.824	4.565	$t = 1.453$ , (D.F.=370.4)
(3a)	Unaccented	188	4.155	4.324	P=0.147

**Table 4: Means and SDs of accented and unaccented data clouds**

Speaker	N	Mean slope (SD)		N	Mean intercept (SD)	
		Accented (2a)	Unaccented(3a)		Accented (2a)	Unaccented (3a)
YO	10	-0.271 (0.052)	-0.173 (0.057)	10	220.221 (10.017)	213.905 (10.565)
KM	10	-0.119 (0.044)	0.005 (0.047)	10	116.050 ( 7.903)	105.055 ( 4.691)
NF	10	-0.381 (0.053)	-0.253 (0.100)	10	266.652 (12.806)	241.182 (16.856)

**Table 5: Statistical test of the data shown in Table 4.**

Speaker	T <sup>2</sup>	F (DF)	P	<i>t</i> (DF) P of slope	<i>t</i> (DF) P of intercept
YO	1.906	16.199 (2,17)	0.000	-4.020 (17.8) 0.001	1.372 (17.9) 0.187
KM	2.013	17.106 (2,17)	0.000	-6.019 (17.9) 0.000	-3.783 (17.9) 0.002
NF	1.906	6.972 (2,17)	0.006	-3.575 (13.6) 0.003	3.805 (16.8) 0.001

linear interpolation between the accentual L of preceding accent (of *da're*) and the beginning of the final rise for question rendition which is usually realized on the last mora. However, since this interpretation is based fully on visual inspection, which can not be free from subjectivity, we must seek for a more reliable quantitative method of analysis. The method adopted here is a line fitting by means of the least-squares principle. This method computes a line which represents the overall trend of given data points by minimizing the prediction error. Specifically, when the prediction error is defined as the distance between the line and a data point on the ordinate, it is equivalent to the computation of a regression line in statistics, which is the case for the present experiment. Figure 4 shows examples of lines fitted to accented and unaccented verb stems respectively. In conducting these and other computations of line fitting, the first data point was assigned to the time value of 10ms in order for the comparison across utterances to be possible. Tables 1, 2 and 3 show the means and the standard deviations (SD) of root-mean-square (RMS) prediction error averaged over the accentedness of the verb stem. The mean difference between accented and unaccented verb stems was not significant for all three speakers. Note also that RMS values are greater for unaccented verb stems than for accented verb stems for two out of three speakers.

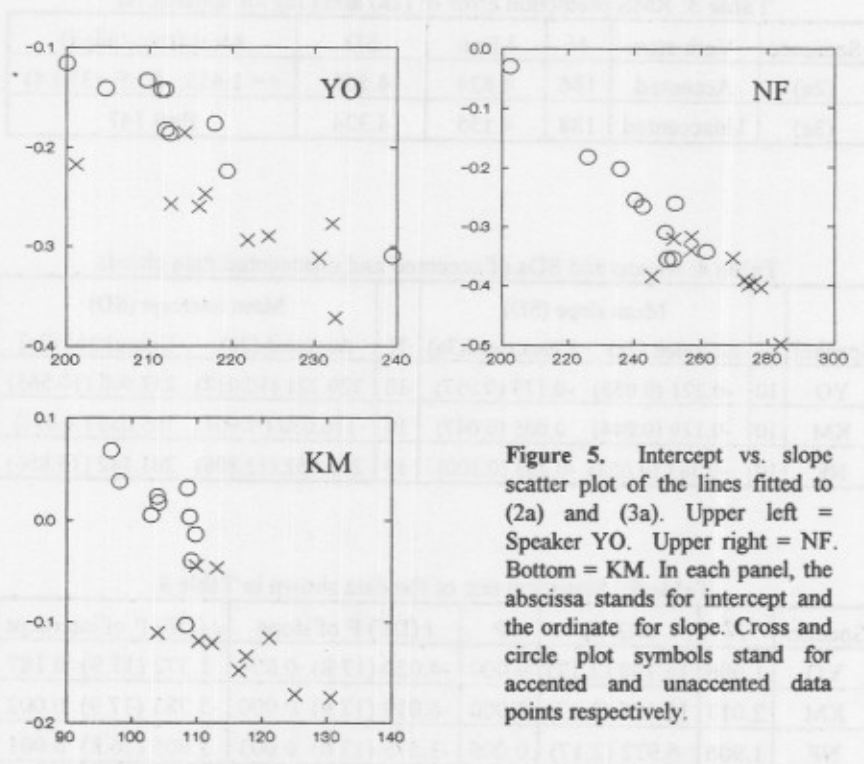


Figure 5. Intercept vs. slope scatter plot of the lines fitted to (2a) and (3a). Upper left = Speaker YO. Upper right = NF. Bottom = KM. In each panel, the abscissa stands for intercept and the ordinate for slope. Cross and circle plot symbols stand for accented and unaccented data points respectively.



I see this as evidence showing that the local pitch fall due to the accent was absent in this context.<sup>7</sup>

The results we have obtained so far seem to show nothing contradictory to the standard assumption of dephrasing. However, the linearity of intonation does not directly assure the occurrence of 'dephrasing', since Tables 1, 2 and 3 do not provide full information concerning the exact shape of f<sub>0</sub> contours. Figure 5 shows scatter plots of slope vs. intercept values of lines fitted to the observed f<sub>0</sub> contours of the three speakers.<sup>8</sup> The data points are classified according to the accentedness of the verb stems. The figure shows clearly that there is a difference between the lines fitted to the accented verb stem and those fitted to the unaccented verb stem. The former lines are characterized by larger intercept values and smaller slope values (in terms of their absolute values). Table 4 shows means and SDs of accented and unaccented data clouds, and the results of statistical tests are summarized in Table 5.

The difference of two-dimensional means between the accented and the unaccented data clouds was tested by Hotelling's T<sup>2</sup> statistic, which follows the F distribution under certain mathematical transformation (Green, 1978).<sup>9</sup> Results of a (univariate) t-test on slope and intercept are shown in Table 5 also. All T<sup>2</sup> values are statistically significant at least at the 0.01 level. Results of t-test confirm the separability of data clouds on each dimension with the sole exception of the intercept of speaker YO. These statistics show clearly that there is a systematic difference of f<sub>0</sub> contour between the two verb stems which is related to the difference of the accentedness of the two lexical items. If we stick to the view that the f<sub>0</sub> contours exemplified by the upper panel of Figure 3 are an instance of dephrasing, it is impossible to provide a linguistic explanation for the observed physical difference.<sup>10</sup>

## 2.2. 'Dephrasing' of auxiliary verbs

The second experiment dealt with the following sentences. Phonological contrasts between (4) and (5) are attributable to the choice of accented or unaccented auxiliary verbs, i.e. /mi'ru/ or /iru/.

- (4a) na'ni-o      no'n-de-mi'ru-n-desu-ka?  
what-OBJ    drink-GER-try-COMP-copula-Q  
'What will (you) try drinking?'

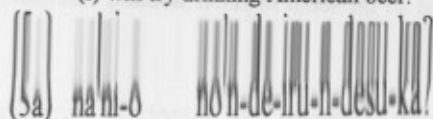
<sup>7</sup> The means and SDs of speaker NF are considerably greater than the values of YO and KM. In fact, some of her utterances of (2a) and (3a) would be likely to fall under Kubozono's category B.

<sup>8</sup> As is known in analytic geometry, the shape of a line can be parameterized perfectly by its slope and intercept.

<sup>9</sup> Hotelling's T<sup>2</sup> is a common diagnostic in multivariate statistical analysis. See Maekawa (1989) for the application of this statistic for the study of vowel merger.

<sup>10</sup> Needless to say, there is a semantic difference between the two verb stems. However, there seems to be no linguistically plausible explanation relating the semantic difference between 'to read' and 'to call' to the observed difference of f<sub>0</sub> contours.

- (4b) no'n-de-mi'ru-no-wa amerika-no bi'iru-desu  
 drink-GER-try-COMP-TOPIC Amerika-GEN beer-copula  
 '(I) will try drinking American beer.'



what-OBJ drink-GER-ing-COMP-copula-Q  
 'What are (you) drinking?'

- (5b) no'n-de-iru-no-wa amerika-no-bi'iru-desu  
 drink-GER-PROG-COMP-TOPIC Amerika-GEN-beer-copula  
 '(I'm) drinking American beer.'

Some native speakers of Tokyo Japanese might feel that the complementizer (COMP) *no* in (4b) and (5b) and the auxiliary verb *iru* in (5a) are accented.<sup>11</sup> Following this intuition, the representation of these sentences would be as follows.

- (4a') na'ni-o no'n-de-mi'ru-n-desu-ka?  
 (4b') no'n-de-mi'ru-no'-waamerika-no bi'iru-desu  
 (5a') na'ni-o no'n-de-iru'-n-desu-ka?  
 (5b') no'n-de-iru-no'-wa amerika-no bi'iru-desu.

This intuition is an important one because it is contradictory to the prediction provided by the standard assumption based on the 'dephrasing' of auxiliary verbs. In any case, supposing that post-focus dephrasing does not take place, the standard assumption predicts the following phrasings.

- (4a'') [na'ni-o]<sub>α</sub> [no'n-de-miru-n-desu-ka]<sub>α</sub>  
 (4b'') [no'n-de-miru-no-wa]<sub>α</sub> [amerika-no bi'iru-desu]<sub>α</sub>  
 (5a'') [na'ni-o]<sub>α</sub> [no'n-de-iru-n-desu-ka]<sub>α</sub>  
 (5b'') [no'n-de-iru-no-wa]<sub>α</sub> [amerika-no bi'iru-desu]<sub>α</sub>

According to this prediction, the f0 contours of the second accental phrases of (4a'') and (5a'') on the one hand, and the f0 contours of the first accental phrases of (4b'') and (5b'') on the other should be exactly the same.

Basically the same procedure of data analysis as that used in the previous experiment was applied for this data set. However, there are difficulties inherent in this data set. For one thing, it is difficult to find a reliable acoustic segmental boundary between /e/ and /i/ of /no'ndeiru/ in (5a,b) since what we observe on spectrograms is continuous transition of vocalic formants from /e/ to /i/. Secondly, and more importantly, the lengths of auxiliary verbs are different in terms of the number of their component segments. Although they are of the same length from a moraic point of view, there is in fact substantial difference in the acoustic duration of /mi'ru/ and /iru/. These points can be critical for the purpose of the present experiment because different acoustic segmentations result directly in different intercept values which play a central role in the data analysis. Since there could be

<sup>11</sup> It is probable that some natives 'feel' an accent before the complementizer /n/ in (4a). In this case we have *na'ni-o no'n-de-mi'ru'-n-desu-ka* instead of (4a') above.

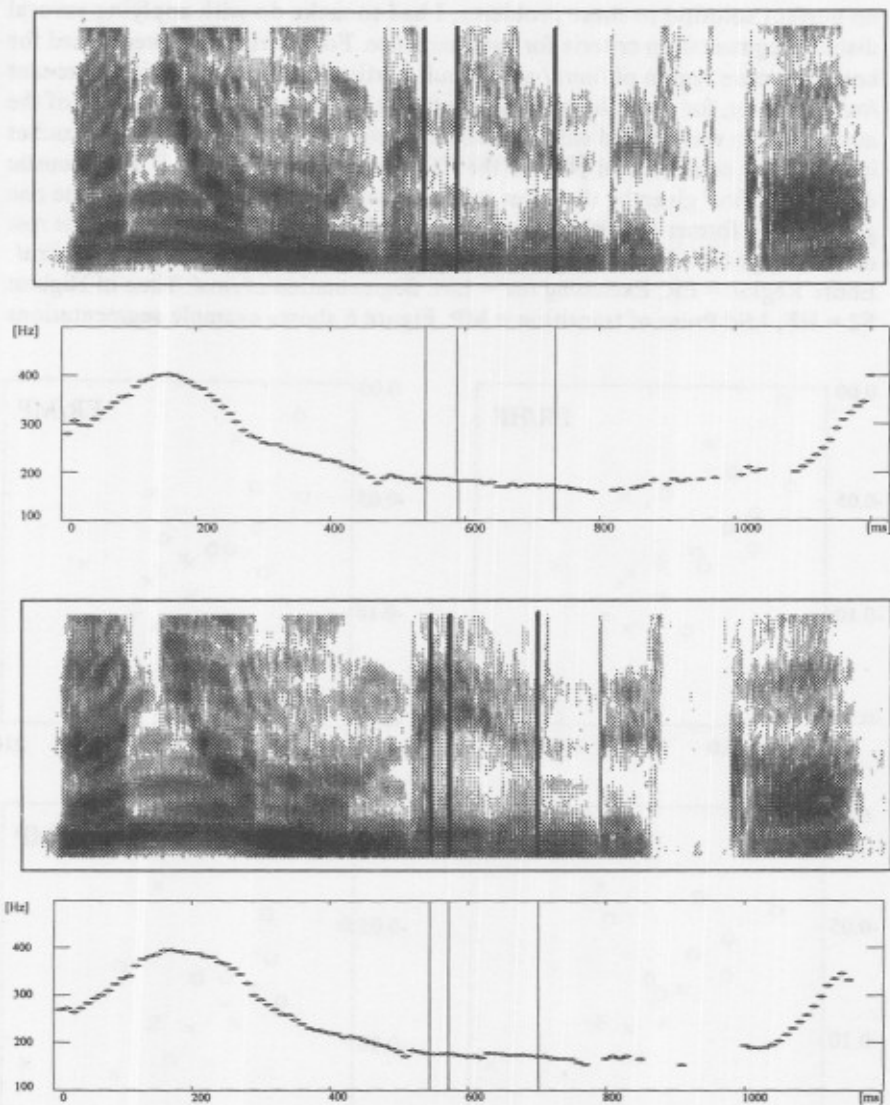
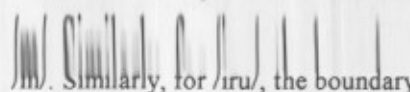


Figure 6. Wide-band spectrograms (4kHz) and f0 contours of (4a) and (5a) by YO. The upper two panels are for (4a). Vertical lines correspond to the beginning of /m/, the end of /m/ and the end of /u/ from left to right. The lower panels are for (5a). Lines correspond to the mid point of F2 transition, the time of the highest F2 and the end of the vowel /u/. These criteria were applied to (4b) and (5b), too.

no perfect solution to these problems, I had to make do with applying several distinct segmentation criteria for each sentence. For /mi'ru/, lines were fitted for both the entire region of /miru/ and its sub-portion omitting the initial consonant



Similarly, for /iru/, the boundary between /e/ of gerundial /de/ and /i/ of the auxiliary verb was defined either as the time point where the F2 transition reaches its maximum or as the mid point of the F2 transition from /e/ to /i/.<sup>12</sup> The acoustic duration of /iru/ given by the latter segmentation is relatively longer than the one given by the former definition. The following abbreviation will be used in the rest of the paper to refer to these segmentation criteria. Segmentation of /mi'ru/: Entire Region = ER, Excluding /m/ = EM. Segmentation of /iru/: Time of Highest F2 = HF, Mid Point of transition = MP. Figure 6 shows example segmentations

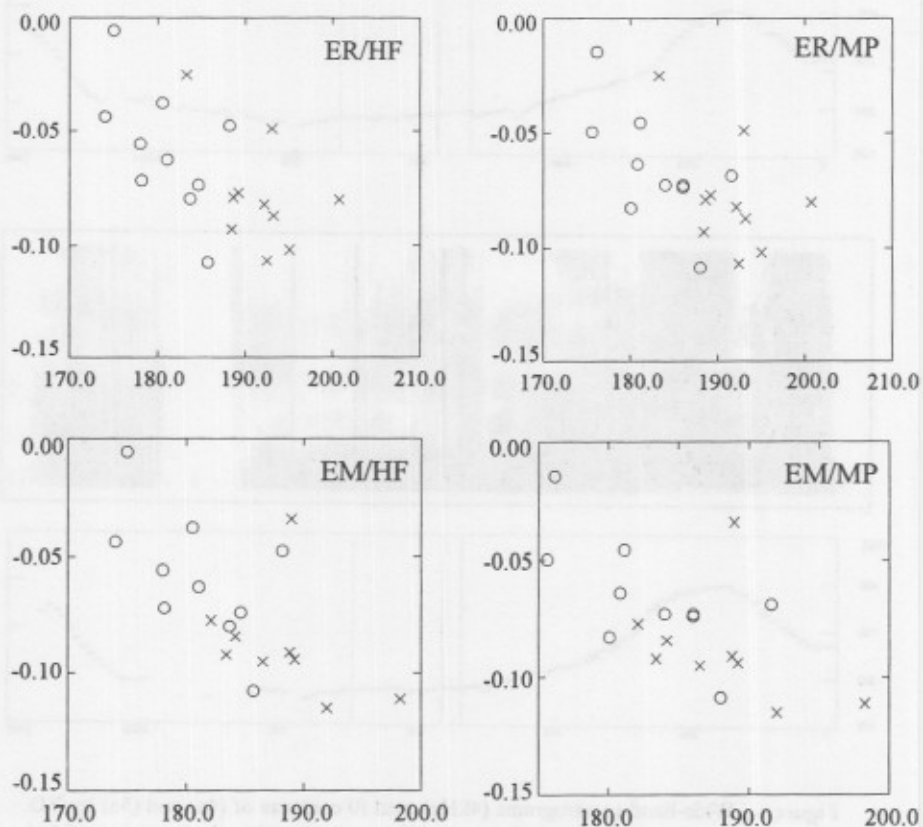


Figure 7. Scatter plot of lines fitted to auxiliary verbs of (4a) and (5a) uttered by YO. The abscissa shows intercept value and the ordinate shows slope. The combination of segmentation criteria is shown on upper right corner of each panel. Cross and circle plot symbols are used for accented and unaccented data respectively.

<sup>12</sup> The right edge of auxiliary verbs could be invariably defined as that of the vowel /u/.

based on these criteria. Note that given the monotonically decreasing  $f_0$  contours, it is predicted that the combination EM/MP disfavors the separation the most in terms of the intercept value since the left edge of /mi'ru/ assigned by criterion EM is relatively late in time (when compared to the edge obtained by ER) and the left edge of /iru/ assigned by MP is relatively early (when compared to HF).

Figure 7 shows the intercept vs. slope scatter plots of lines fitted to the auxiliary verbs in (4a) and (5a) uttered by speaker YO. Figure 8 shows the same for the auxiliary verbs in (4b) and (5b) uttered by the same speaker. Each panel in the figure stands for a specific combination of segmentation criteria. As can be seen from the figure, the separability of accented and unaccented data clouds differs according to the combinations of segmentation criteria. The same analysis was carried out for the utterances of speaker KM. As for speaker NF, only the auxiliary verbs in (4a) and (5a) were analyzed since in her speech phonetic realization of the auxiliary verbs in (4b) and (5b) showed considerable amount of variation through Kubozono's types A to C.

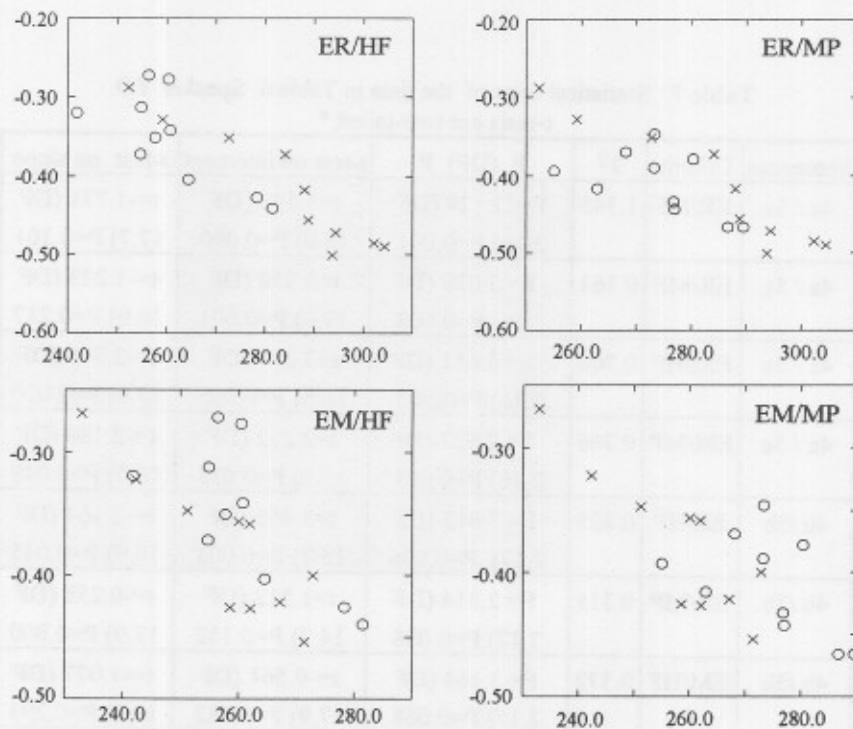


Figure 8. Scatter plots of lines fitted to the auxiliary verbs of (4b) and (5b) uttered by YO. Same manner of plotting as in Figure 7.

**Table 6:** Statistics of auxiliary verbs in (4a,b) and (5a,b) by Speaker YO.

Sentence	Aux. verb	Criterion	N	Mean (SD) of intercept	Mean (SD) of slope
4a	/ mi'ru /	ER	9	192.540 (3.838)	-0.084 (0.017)
4a	/ mi'ru /	EM	9	188.163 (4.957)	-0.088 (0.024)
5a	/ iru /	HF	10	180.900 (4.670)	-0.058 (0.028)
5a	/ iru /	MP	10	182.949 (5.128)	-0.065 (0.025)
4b	/ mi'ru /	ER	10	283.991 (17.447)	-0.416 (0.076)
4b	/ mi'ru /	EM	10	258.049 (12.645)	-0.378 (0.058)
5b	/ iru /	HF	10	261.094 (11.609)	-0.350 (0.059)
5b	/ iru /	MP	10	274.276 (10.409)	-0.409 (0.041)

**Table 7:** Statistical tests of the data in Table 6. Speaker YO.  
t-tests are two-tailed.\*

Sentences	Criteria	T <sup>2</sup>	F (DF) P	t-test on intercept	t-test on slope
4a / 5a	ER/HF	1.545	F=13.129 (DF 2,16) P=0.001	t=5.140 (DF 18.0) P=0.000	t=-1.731 (DF 17.7) P=0.101
4a / 5a	ER/MP	0.363	F= 3.089 (DF 2,16) P=0.003	t=3.958 (DF 17.8) P=0.001	t=-1.223 (DF 18.0) P=0.237
4a / 5a	EM/HF	0.706	F= 5.651 (DF 2,16) P=0.005	t=3.227 (DF 16.5) P=0.005	t=-2.576 (DF 17.0) P=0.020
4a / 5a	EM/MP	0.366	F= 2.927 (DF 2,16) P=0.083	t=2.252 (DF 16.9) P=0.038	t=-2.188 (DF 16.9) P=0.049
4b / 5b	ER/HF	0.825	F= 7.013 (DF 2,17) P=0.006	t=3.455 (DF 15.7) P=0.003	t=-2.164 (DF 16.9) P=0.045
4b / 5b	ER/MP	0.311	F= 2.814 (DF 2,17) P=0.088	t=1.512 (DF 14.7) P=0.152	t=-0.258 (DF 13.9) P=0.800
4b / 5b	EM/HF	0.372	F= 3.164 (DF 2,17) P=0.068	t=-0.561 (DF 17.9) P=0.582	t=-1.077 (DF 18.0) P=0.295
4b / 5b	EM/MP	0.658	F= 5.597 (DF 2,17) P=0.014	t=-3.133 (DF 17.4) P=0.006	t=1.374 (DF 16.4) P=0.188

\*Shaded cells are significant at least at 0.05 level.

**Table 8:** Statistics of auxiliary verbs in (4a,b) and (5a,b) by speaker KM.

Sentence	Aux. verb	Criterion	N	Mean (SD) of intercept	Mean (SD) of slope
4a	/ mi'ru /	ER	10	104.632 ( 4.485)	-0.119 (0.025)
4a	/ mi'ru /	EM	10	97.265 ( 4.342)	-0.102 (0.028)
5a	/ iru /	HF	10	91.958 ( 2.618)	-0.090 (0.033)
5a	/ iru /	MP	10	97.342 ( 2.765)	-0.110 (0.028)
4b	/ mi'ru /	ER	10	148.322 ( 8.355)	-0.240 (0.043)
4b	/ mi'ru /	EM	10	130.282 ( 7.311)	-0.182 (0.055)
5b	/ iru /	HF	10	133.108 ( 8.267)	-0.213 (0.072)
5b	/ iru /	MP	10	145.295 (10.401)	-0.266 (0.064)

**Table 9:** Statistical tests of the data in Table 8. Speaker KM.

t-tests are two-tailed.\*

Sentences	Criteria	T <sup>2</sup>	F (DF) P	t-test on intercept	t-test on slope
4a / 5a	ER/HF	3.313	F=28.157 (DF 2,17) P=0.000	t=7.717 (DF 14.5) P=0.000	t=-2.153 (DF 16.8) P=0.046
4a / 5a	ER/MP	1.092	F=9.281 (DF 2,17) P=0.002	t=4.357 (DF 15.0) P=0.001	t=-0.764 (DF 17.8) P=0.455
4a / 5a	EM/HF	0.614	F=5.219 (DF 2,17) P=0.017	t=3.310 (DF 14.8) P=0.005	t=-0.882 (DF 17.4) P=0.390
4a / 5a	EM/MP	0.019	F=0.159 (DF 2,17) P=0.854	t=-0.854 (DF 15.3) P=0.963	t=0.572 (DF 18.0) P=0.574
4b / 5b	ER/HF	1.723	F=14.645 (DF 2,17) P=0.000	t=4.093 (DF 18.0) P=0.001	t=-1.039 (DF 14.6) P=0.316
4b / 5b	ER/MP	0.516	F=4.383 (DF 2,17) P=0.029	t=0.718 (DF 17.2) P=0.483	t=1.075 (DF 15.8) P=0.298
4b / 5b	EM/HF	0.062	F=0.528 (DF 2,17) P=0.599	t=0.810 (DF 17.7) P=0.429	t=0.810 (DF 16.8) P=0.305
4b / 5b	EM/MP	0.783	F=6.656 (DF 2,17) P=0.007	t=3.734 (DF 16.1) P=0.002	t=-3.162 (DF 17.7) P=0.005

\*Shaded cells are significant at least at 0.05 level.

**Table 10:** Statistics of auxiliary verbs in (4b) and (5b) by speaker NF.

Sentence	Aux. verb	Criterion	N	Mean (SD) of intercept	Mean (SD) of slope
4b	/ mi'ru /	ER	10	347.872 (44.234)	-0.665 (0.131)
4b	/ mi'ru /	EM	10	310.807 (42.212)	-0.626 (0.161)
5b	/ iru /	HF	10	327.809 (39.887)	-0.739 (0.207)
5b	/ iru /	MP	10	339.522 (53.338)	-0.749 (0.194)

**Table 11:** Statistical tests of the data in Table 10. Speaker NF.  
t-tests are two-tailed.\*

Sentences	Criteria	T <sup>2</sup>	F (DF) P	t-test on intercept	t-test on slope
4b / 5b	ER/HF	0.979	F=8.318 (DF 2,17) P=0.003	t=1.065 (DF 17.8) P=0.301	t=0.952 (DF 15.2) P=0.356
4b / 5b	ER/MP	0.999	F=8.490 (DF 2,17) P=0.003	t=0.381 (DF 17.4) P=0.708	t=1.128 (DF 15.8) P=0.276
4b / 5b	EM/HF	0.145	F=1.231 (DF 2,17) P=0.317	t=-0.926 (DF 17.9) P=0.367	t=1.354 (DF 17.0) P=0.194
4b / 5b	EM/MP	0.141	F=1.198 (DF 2,17) P=0.326	t=-1.335 (DF 17.1) P=0.199	t=1.531 (DF 17.4) P=0.144

\* Shaded cells are significant at least at 0.05 level.

Tables 6 and 7 show the results of statistical analysis of the utterances of speaker YO. In six out of eight combinations of criteria, statistical significance of at least at the 0.05 level is obtained at least in more than two statistical tests, the remaining two cases (ER/MP and EM/HF of (4b)/(5b)) are significant at the level of 0.1 in terms of their T<sup>2</sup> values. However, it should be pointed out that in the case of EM/MP of (4b)/(5b), the mean intercept value of the unaccented data is greater than that of the accented data. This is the reverse of what was observed in the previous experiment for all three speakers. This reversal could happen when the left edge of /iru/ assigned by the criterion MP is too early on the time axis and is penetrating into the time region of the accentual fall caused by the accent of *no'nde*. In fact, the temporal location of the accentual L of /no'nde/ seems to coincide sometimes with the beginning of /t/ of /iru/.

The result of speaker KM shown in Tables 8 and 9 is basically the same as that for YO. Six out of eight combinations were significant at least at the 0.05 level in terms of their T<sup>2</sup> values. The remaining cases, EM/MP of (4a)/(5a) and EM/HF of (4b)/(5b), are not significant in any of the three tests at any level, however. The segmentation by the criteria EM/MP causes the same reversal of parameter values as in YO's utterances, but here it is not only the relation between (4b) and (5b) but also that of (4a) and (5a) that is reversed. The result of speaker NF is shown in



Tables 10 and 11. Here again, the combination of EM/MR caused the same reversal, although in her speech the measurement provided by this combination did not show significance at any level.

### 3. Discussion and concluding remarks

Concerning the 'dephrasing' in post-focus position, there seems to be almost no room for controversy about the detectability of the accentedness of seemingly 'dephrased' accentual phrases. Nearly perfect separability of data clouds on the slope vs. intercept plane shown in Figure 5 provides strong support for the interpretation that the  $f_0$  contour of accented accentual phrase like the one shown in the upper panel of Figure 3 above is not an instance of dephrasing. Therefore the two contours of Figure 3 can not be identical from a phonological point of view. No matter how its physical manifestation differs from the canonical form, the accent of the seemingly 'dephrased' accented accentual phrase is still there. In this sense, it is appropriate to call this kind of contour a *degenerate* accentual phrase and distinguish it from the 'true' dephrasing in which the accentual contrast can not be detected anymore. In this respect, the intuitive judgment that the accent in the predicate of sentence (1c) can be perceived even when the predicate is associated with linear  $f_0$  contour is not surprising at all, since the linear synthetic intonation used in Maekawa (1991) had a relatively high intercept value and showed relatively steep declination. Perception of the accent in the linear  $f_0$  contour should not be treated as a perceptual illusion due to speakers' internalized knowledge of the accentedness.

When compared to the clarity in the result of the first experiment, the result of the second experiment leaves much room for controversy. However, the results of statistical tests showed arguably that it was possible to detect the accentedness of auxiliary verbs at least in the speech of speaker YO, and therefore this speaker showed no 'dephrasing' of auxiliary verbs. As for the other two speakers, although it is possible to reject the 'dephrasing' view based on the results of statistical tests which showed significant differences in half of the case examined, it would be safer to eschew any decisive conclusion at this time. (The data set of an experiment currently underway involves sentences similar to (4) and (5) in which the unaccented auxiliary verb *yaru* ('do something for someone') is used instead of *mi'ru* and *iru*. Comparison of *mi'ru* and *yaru* would contribute to the clarification of the problem since the segmental lengths of the two auxiliary verbs are the same. But in this case, we must face the difference in vocalic intrinsic pitch caused by the difference of the vowel height of the auxiliary verbs; this was why I analyzed the pair *mi'ru* and *iru* in the previous experiment.)

To sum up, although the experimental evidence presented in this paper should be reinforced by analyzing a wider range of data and more informants, it suggests strongly the view that contours hitherto regarded as 'dephrased' are not dephrased. At the same time, the data presented in this paper show convincingly that it is risky to rely solely upon the visual, hence subjective, inspection of  $f_0$  data, especially the visual detection of the peak features.

Finally, I would like to point out two issues left unresolved by the present study. First, it remains as an open question how to treat the 'dephrasing' of unaccented accentual phrases, i.e. the treatment of the  $f_0$  contours like the ones shown in the lower panels of Figure 3 or Figure 6.<sup>13</sup> From a theoretical point of view, to regard these contours as dephrased causes less problem than the dephrasing of an accented accentual phrase because the dephrasing of an unaccented accentual phrase does not involve accent deletion. Intuitively, however, I feel that I perceive the existence of an accentual phrase boundary when the slope value of the quasi-linear—hence seemingly 'dephrased'—unaccented accentual phrase is very close to or higher than zero. In the course of the present study, I have tried to screen out the true cases of dephrasing by examining the normal probability plot of extracted slope values of sentences (3a,b) uttered by KM; his slope values seemed to be split into two distinct distributions on both sides of the zero point. But I could not arrive at any clear conclusion due mainly to the limitation of the number of data. I think, however, that this method is worth further effort.

The second problem is a more important one. At present it is unclear whether degenerate accented accentual phrases have the same phonological representation as their fully realized counterparts. It seems to me that there is no *a priori* reason to suppose that degeneration of accentual phrase is a pure phonetic process as supposed by Kubozono. As far as I am aware of, one way to examine this problem involves the quantitative comparison between the effect of downstep caused by a fully realized accentual phrase and the one caused by a degenerate accentual phrase. Fragmentary analysis of relevant material contained in the current data set suggests that there is difference between the two types of accentual phrases. This problem should be discussed in a separate paper.

### References

- Beckman, M. and Pierrehumbert, J. (1986) Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255-309.
- Fujimura, O. (1991) Remarks on phrasing and prosodic attachment. In *Interdisciplinary approaches to language* (C. Georgopoulos & R. Ishihara, editors), pp.207- 216, Netherlands, Kluwer Academic Publishers.
- Green P.E.(1978) *Analyzing multivariate data*. Hinsdale, IL: The Dryden Press.
- Kubozono, H. (1993) *The organization of Japanese prosody*. Tokyo: Kuroshio Publishers.
- Maekawa, K. (1989) Statistical tests for the study of vowel merger. *Quantitative Linguistics*, 39, 200-219, Bochm.
- Maekawa, K. (1991) Perception of intonational characteristics of WH and NON-WH questions in Tokyo Japanese. In *Proceedings 12th International Congress on Phonetic Sciences*, Aix-en-Provence, 4, 202-205.
- McCawley, J. (1968) *The phonological component of a grammar of Japanese*. The Hague: Mouton.

<sup>13</sup> This problem was pointed out to me by Jennifer Venditti (personal communication).

- Pierrehumbert, J. (1980) *The phonology and phonetics of English intonation*. Ph.D. dissertation, MIT.
- Pierrehumbert J. and M. Beckman, (1988). *Japanese tone structure*. Linguistic Inquiry Monograph Series, 15, Cambridge, MA: MIT Press.
- Poser, W. (1984) *The phonetics and phonology of tone and intonation in Japanese*. Ph.D. dissertation, MIT.
- Wada, M. (1969) Ji-no akusento (Accent of particles), *Kokugokenkyuu*, 29.

## Effects of Prosodic Position and Tonal Context on Taiwanese Tones\*

Shu-hui Peng  
peng@ling.ohio-state.edu

**Abstract:** The present study investigates the effects of prosodic position and following tone on duration and fundamental frequency of voicing (F0) pattern for the Taiwanese tones. The results showed that Taiwanese tones were acoustically influenced by prosodic position and to a certain extent by tonal context. Prosodic position had strong effects on both the F0 and duration of tones. Final lengthening and final lowering were found in the utterance-final and to a somewhat lesser extent in phrase-final positions. Pitch range was substantially affected by prosodic position, but in general, the tone shape was not changed. Tone sandhi in Taiwanese was not only defined by syntactic phrase but also by prosodic phrase. Anticipatory tonal coarticulation affected the F0 offsets of Taiwanese tones without affecting syllable duration. Assimilation occurred between contour tones and the following tone. Dissimilation was found between the high-level tone and the following tone. Pitch range in general was not affected by tonal context; however, some subject-dependent variation was found.

### Introduction

It has been found that durations of syllables and segments in sentences are conditioned by prosodic position. For example, syllables at the end of utterances are longer than those in non-utterance-final positions (Fonagy & Magdics, 1960). This phenomena is known as pre-pausal lengthening of syllables. A study of vowel duration in a connected discourse (Klatt, 1975) indicates that a vowel is longer not only at the end of an utterance but also at the end of a prosodic phrase. Klatt also found that the lengthening of vowels at the end of utterances is not greater than that in the end of a phrase. A descriptive model of segment duration by Lindblom and Rapp (1973) suggests that segments in the final syllables of a phrase are longer than in non-final syllables, since segments in the middle of a phrase are temporally more compressed than those in other positions of the phrase.

The intonation of utterances is shaped by prosodic position. Liberman & Pierrehumbert (1984) found that the fundamental frequency (F0) of the last accent of an utterance is lower than that of the non-final accent in the same position in a longer utterance. This effect of phrasal position on the F0 of utterances is known as final-lowering. Final-lowering also exists in tone languages. A study by Shih (1988) comparing the same Mandarin tone in different positions of an utterance showed that the F0 of the tone in utterance-final position was lower than in

---

\* I thank Mary Beckman, Marjorie Chan and Keith Johnson for their helpful comments. I am also grateful to Sun-Ah Jun and Jennifer Venditti for their help on an earlier version of this paper and Sook-Hyang Lee for her assistance with the analysis of variance.

utterance-initial and utterance-medial positions. It is in the initial position that the target tone has the highest value of F0.

For tone languages, in addition to prosodic position, tonal context is another factor that may affect the F0 of tones. In some tone languages, there are regular phonological rules (tone sandhi) for tones which change one tone into another tone. However, in all tone languages some tonal coarticulation is expected. Tonal coarticulation, namely the phonetic influence of one tone on another tone, occurs between adjacent tones due to their close specification. According to the direction in which coarticulation applies, tonal coarticulation can be classified into two different types: anticipatory and perseveratory. Anticipatory coarticulation occurs when the F0 pattern on one syllable is influenced by the tonal features of the syllable following it. Perseveratory coarticulation is the carry-over of the tonal features of a preceding syllable.

In Mandarin (Shih, 1988; Shen, 1990), tonal coarticulation is bi-directional: anticipatory and perseveratory. Shih (1988) has analyzed tones according to pitch levels, and thus distinctive tones in Mandarin are represented with different tonal targets which correspond to the pitch levels of the tones. For instance, the rising tone (tone 2) is represented as mid-high (MH). The relative F0 values and timing of tonal targets might be shifted due to tonal coarticulation. Shih (1988) found that the rising tone (tone 2) ends higher when the following target is low, but ended lower when before high tonal target. The rising tone followed by a high target ends lower due to the shift of the final high target of the rising tone to the following high target. Conversely, the rising tone followed by a low target was not affected by the following low target.

Moreover, a later Mandarin study by Shen (1990) found that in addition to onsets and offsets of tones, the overall tonal height (pitch ranges) of tone contours are also affected by neighboring tones. The overall pitch ranges of tones following high-offset tones are higher than those of tones following low-offset tones. However, the directions of tone contours were not shifted by neighboring tones. That is, in general, tone shape was constant while onset and offset values and the overall tone height vary depending on neighboring tones. It would be interesting to see whether these effects of prosodic position and tonal context are language-specific phonetic processes in Mandarin or whether they also exist in a related tone language like Taiwanese.

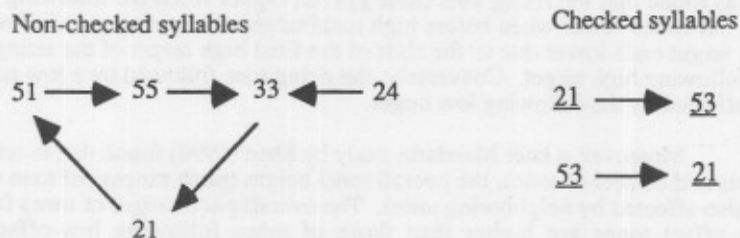
The Taiwanese tonal system is different from the Mandarin one in two ways. First, there are more phonologically distinctive tones in Taiwanese. Second, in Taiwanese, tone sandhi is not conditioned by tonal context but only by prosodic position; all but the last syllables of tonal phrases change to sandhi forms. Given these differences, we might expect a different pattern of tonal coarticulation in Taiwanese.

The purpose of the present study is to examine the effect of prosodic context and tonal context on Taiwanese tones using the measurements of duration and F0. There are seven phonologically distinctive tones in Taiwanese Amoy. Two of them are in checked syllables—i.e. syllables which end with stop consonants. Five of them are in non-checked syllables—syllables which do not end with stop consonants. Descriptions and transcriptions of the Taiwanese tones using the traditional 5-point numeric scale (Chao, 1930) are listed in Table 1.

**Table 1.** Qualitative descriptions and numeric scales of Taiwanese tones  
(Checked tones are underlined.)

Description	Numeric scale	example
High-level	55	/tu <sup>55</sup> / "to push"
Mid-level	33	/tu <sup>33</sup> / "to consider"
Mid-rising	24	/tu <sup>24</sup> / "to be like"
Mid-falling	21	/tu <sup>21</sup> / "to scrape"
High-falling	51	/tu <sup>51</sup> / "woman"
Mid-falling checked	<u>21</u>	/tu <sup><u>21</u></sup> / "to rub away"
High-falling checked	<u>53</u>	/tu <sup><u>53</u></sup> / "rule"

**Table 2.** Taiwanese tone sandhi rules. One tone changes into another when followed by another tone in the same tonal phrase.



Examples:

before tone sandhi	after tone sandhi	meaning
/tsa <sup>51</sup> / "early" /tɿ <sup>21</sup> / "meal"	[tsa <sup>55</sup> tɿ <sup>21</sup> ]	"breakfast"
/tiɔŋ <sup>55</sup> / "middle" /tau <sup>21</sup> / "daytime"	[tiɔŋ <sup>33</sup> tau <sup>21</sup> ]	"at noon"
/aŋ <sup>24</sup> / "red" /hwe <sup>55</sup> / "flower"	[aŋ <sup>33</sup> hwe <sup>55</sup> ]	"red flower"
/tau <sup>33</sup> / "bean" /liŋ <sup>55</sup> / "milk"	[tau <sup>21</sup> liŋ <sup>55</sup> ]	"soy bean milk"
/pai <sup>21</sup> / "to worship" /gɔ <sup>33</sup> / "five"	[pai <sup>51</sup> gɔ <sup>33</sup> ]	"Friday"
/sit <sup>21</sup> / "to lose" /le <sup>51</sup> / "politeness"	[sit <sup>53</sup> le <sup>51</sup> ]	"Excuse me!"
/lut <sup>53</sup> / "law" /su <sup>55</sup> / "expert"	[lut <sup>21</sup> su <sup>55</sup> ]	"lawyer"

Like many other similar tone languages, Taiwanese has tone sandhi rules. Each Taiwanese tone undergoes tone sandhi when followed by another syllable in the same tonal phrase. The changes of tones are systematic and predictable and are conditioned by prosodic position, but not by the presence of a specific tone as

Mandarin tone sandhi rules. Each tone is changed into another tone regardless of the tone quality of the tone following it (Cheng, 1968). These changes form a Tone Sandhi Chain, as shown in Table 2. The tone in final position of a tonal phrase or followed by a final neutral tone stays unchanged, i.e. does not undergo tone sandhi. Tones preceding the final syllables of tonal phrases are changed by tone sandhi rules. For example, /bi<sup>51</sup>/ "beautiful" and /lu<sup>51</sup>/ "woman" become /bi<sup>55</sup> lu<sup>51</sup>/ when they form a noun phrase "beautiful woman".

It has been demonstrated that tonal phrasing is closely related to syntactic phrasing in Taiwanese (Cheng, 1968; 1973; Chen, 1987). Syllables in phrases and sentences are grouped into tonal phrases based in large part on syntactic structure. However, there are some exceptions to the formation of tonal phrases due to rhythmic structure (Hsiao, 1990). For example, /taŋ<sup>55</sup>/ "east", /sai<sup>55</sup>/ "west", /lam<sup>24</sup>/ "south" and /pak<sup>21</sup>/ "north" become [taŋ<sup>33</sup> sai<sup>55</sup> lam<sup>33</sup> pak<sup>21</sup>], when they are read in a sequence. The first two syllables and the last two syllables form separate rhythmic feet. Only the first tone of each foot is changed. In addition to these phonological changes created by tone sandhi, phonetic variations caused by tonal context and prosodic position are also expected.

The present study investigated five distinctive Taiwanese tones in different prosodic positions: phrase-initial, phrase-medial, phrase-final and utterance-final as well as the coarticulatory effects of the following tone. It reports results from two pilot studies using only one talker and a larger main study using four talkers. The first pilot study was designed to see whether Taiwanese high-level and high-falling tones are influenced by prosodic position and the following tone. Also, it investigated whether the influence on different target tones follows the same overall pattern. Results indicated that prosodic position affected both the F0 and duration of the high-level tone, but only the F0 of the high-falling tone. The F0 and duration of the high-level tone also varied depending on tonal context. A dissimilation-like effect on F0 was found between the target tone and the following tone. In order to examine the effect of following tone even further, the second pilot study focused on the high-level tone followed by five distinctive tones in phrase-medial and phrase-final positions. Very similar results were found in the measurements of duration and F0.

The main study then included all five Taiwanese tones in non-checked syllables followed by tones with various F0 onset values: high-level tone, mid-level tone or mid-falling tone in the same four prosodic positions used in the first pilot study. The results showed that prosodic position affected the F0 and duration of Taiwanese tones more strongly than did following tone. Anticipatory tonal coarticulation was found. These findings contradict the results of Lin's (1988) study in which Taiwanese tonal coarticulation was found to be only perseveratory coarticulation.

## First Pilot Study

### Method

#### Materials

The corpus used in this experiment is shown in Appendix 1. The target syllable which was either high-level tone [kɔ<sup>55</sup>] or high-falling tone [kɔ<sup>51</sup>] was

placed in four different prosodic positions: tone-phrase-initial, tone-phrase-medial, tone-phrase-final and utterance-final. When the high-level target tone was not utterance-final, it was followed by a syllable which either had high-level tone or mid-level tone except that in phrase-initial position it did not occur before mid-level tone. The syllable following the high target tone started with a voiceless unaspirated stop consonant (except for one case in which it was voiced). The high-falling target tone was followed by a syllable with high-level tone and starting with a voiceless unaspirated stop consonant.

Each target tone was embedded in three five-syllable sentences and one three-syllable phrase as shown in Table 3. Each sentence contained two tonal phrases. The phrasal boundary was after the second syllable in the first of the three sentences. Here the target tone was in the initial position of the second phrase (phrase-initial case). The phrasal boundary was after the third syllable in the other sentences. The target tone was at the medial position (phrase-medial case) or the final position of the first phrase (phrase-final case). In addition, a three-syllable utterance was used. The target tone was at the final position of this utterance (utterance-final case).

**Table 3.** Schematization of corpus 1 utterances for different prosodic positions with the target tone (syllable) underlined. Square brackets indicate the boundaries of tonal phrases.

Prosodic positions	schematized utterances
phrase-initial	[σ σ] [σ̲ σ σ]
phrase-medial	[σ σ̲ σ] [σ σ]
phrase-final	[σ σ σ̲] [σ σ]
utterance-final	[σ σ σ̲]

There were eleven utterances in total. Each of them was read five times. Therefore, there were fifty-five tokens. Nine filler tokens were used. Tokens were arranged in random order.

#### *Subject*

One female native Taiwanese speaker (HS) who also speaks Mandarin and English fluently participated the present experiment.

#### *Recording & Acoustic Measurements*

The subject was asked to read the tokens in a normal, fluent speaking style. The productions of the subject were recorded using a TEAC V-427C stereo cassette deck in a sound-proof booth in the Ohio State University Linguistics Laboratory.

Durations and fundamental frequencies of target tones were measured. Durations of target tones were analyzed with wide-band spectrograms using the Kay DSP 5500. The duration of each target tone was measured from the release of



the initial stop consonant to the offset of the vowel. In addition, each sentence or phrase was digitized and its fundamental frequency contours analyzed with the *waves*™ signal editor (Entropic Research Laboratory, INC., 1993) on a Sun Sparcstation. The F0 of each high target tone (55) was measured at 3 points along the frequency contour for the syllable: onset, mid-point and offset of the F0 contour. For each high-falling target tone (51), only onset and offset of the F0 contour were measured. Mean duration and F0s of each target tone were calculated. Because the tone preceding the target tone was not controlled in the present experiment, all onset values of F0 contours were only used to show the general pitch range of pitch contours. The test sentence in which the high target tone at the phrase-initial position is followed by the mid-level tone was deleted due to a design error.

## Results

A two-way analysis of variance (ANOVA) (prosodic position x tonal context) was done for each target tone. The effect of prosodic position on duration of the high target tone is significant,  $[F(3, 26) = 70.02, p < 0.01]$ . Figure 1 shows the overall mean measurements of the high-level target tone pooled over the two following tone contexts at phrase-medial, phrase-final and utterance-final positions. Three F0 values are plotted: the point at time zero representing the onset F0 of the syllable, the next point representing the F0 in the middle of the syllable and the last point representing the F0 at the end of the syllable

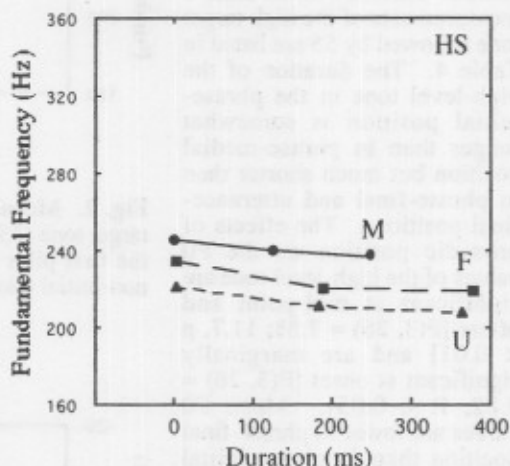


Fig. 1. Mean durations and F0s of the high target tone (55) in different prosodic positions in the first pilot study. Means pooled over the two tonal contexts. M = phrase-medial, F = phrase-final, U = utterance-final.

Table 4. Mean durations and F0s of the high target tone followed by high-level tone (55) in different prosodic positions. F0 is measured at three different points along the pitch contour: onset, mid-point and offset.

prosodic position	duration (ms)	F0 (Hz)		
		onset	mid-point	offset
phrase-initial	254.4	242.92	232.82	233.3
phrase-medial	230.8	230.88	231.04	236
phrase-final	392.4	228.1	208.2	207.2
utterance-final	362.2	220.8	210.76	207.74

(and the abscissa showing mean syllable duration). As the figure shows, durations of the high-level tone in phrase-final and

utterance-final positions are longer than in phrase-medial position. Since the target syllable in phrase-initial position occurred only before 55, we can only compare all four prosodic contexts only in that tonal context. Mean durations and F0 measurements of the high target tone followed by 55 are listed in Table 4. The duration of the high-level tone in the phrase-initial position is somewhat longer than in phrase-medial position but much shorter than in phrase-final and utterance-final positions. The effects of prosodic position on the F0 values of the high-level tone are significant at mid-point and offset [ $F(3, 26) = 7.88; 11.7, p < 0.01$ ] and are marginally significant at onset [ $F(3, 26) = 3.82, P < 0.05$ ]. Mean F0 values are lower in phrase-final position than in phrase-initial and phrase-medial positions and lower still in utterance-final position.

Figure 2 shows the mean measurements pooled over prosodic positions common to the two tonal contexts. The duration of the high-level tone does not vary according to following tone. However, the effect of following tone on the height of F0 is significant at all measurement points [ $F(1, 20) = 18.78; 32.81; 14.51, p < 0.01$ ]. A dissimilation-like effect of the following tone level was found: F0 of the high-level tone is higher when followed by mid-level tone than when followed by another high-level tone.

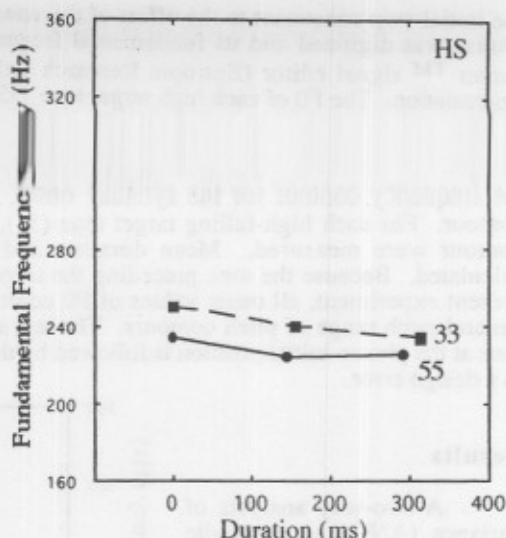


Fig. 2. Mean durations and F0s of the high target tone (55) followed by different tones in the first pilot study. Means pooled over all non-initial tokens.

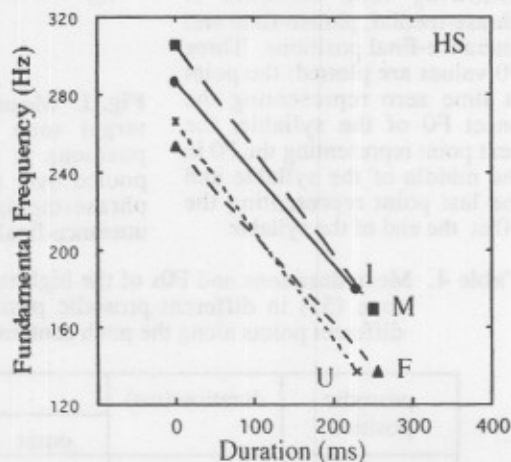


Fig. 3. Mean durations and F0s of the high-falling tone (51) in different prosodic positions. I = phrase-initial, M = phrase-medial, F = phrase-final, U = utterance-final.

As shown by Figure 3, the duration of the high-falling tone is not significantly changed by prosodic position, [ $F(3, 16) = 2.96, p > 0.05$ ], but the pitch range of F0 is, [ $F(3, 16) = 22.6; 15.54, p < 0.01$ ]. Mean values are lower in the phrase-final and utterance-final positions than in the phrase-initial and phrase-medial positions.

## Second Pilot Study

The aim of this pilot experiment is to examine the effects of the following tone and phrasal position even further by adding more tonal contexts to the target tone and focusing on two phrasal positions: phrase-medial and phrase-final.

### Method

#### Materials

The corpus used in this pilot experiment is listed in Appendix 2. A syllable with high-level tone [po<sup>55</sup>] was the target syllable. It was placed either in phrase-medial position or in the phrase-final position, and was followed by any of the five tones found in open syllables. However, there was no final target in the context of the rising tone, because here the following syllable began the next phrase and the rising tone as a surface form only appears in phrase-final position because of its unique status in tone sandhi cycle as shown in Table 2. The prosodic structure of the sentences was the same as those in the first pilot study (shown in Table 3). Thus there were nine different sentences in total: five for phrase-medial position and four for phrase-final position. Each of the sentences was read six times. Eighteen filler sentences were added to the fifty-four test sentences. Tokens were read in random order.

#### Subject

The same subject who participated the first pilot study also participated this experiment.

#### Recording and Acoustic Measurements

The same methods of recording and measurement used in the first pilot study were followed. Only five of the six repetitions of each sentence and phrase which showed complete pitch contours were selected to be measured for duration and fundamental frequency.

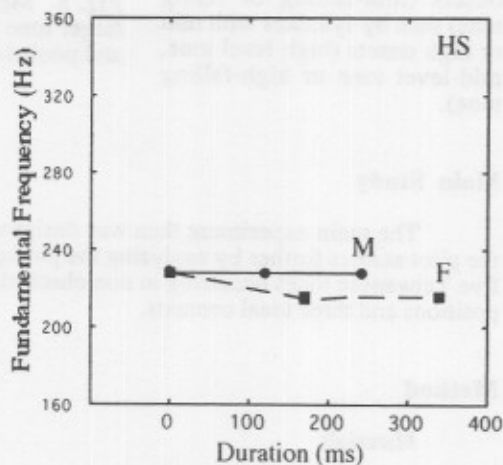


Fig. 4. Mean durations and F0s of the high target tone (55) in the phrase-medial position (M) versus the phrase-final position (F) pooled over tonal context.

## Results

As in the first pilot study, the effect of prosodic position on the duration of the high-level tone is bigger than its effect on fundamental frequency (Figure 4). The mean duration of the high-level tone in phrase-final position is significantly

longer than at phrase-medial position, [ $F(1, 32) = 125.7, p < 0.01$ ]. The effect of prosodic position on F0 of the high-level tone is significant at both the mid-points, [ $F(1, 32) = 9.16, p < 0.01$ ] and the offsets of the F0 contours, [ $F(1, 32) = 12.33, p < 0.01$ ].

Figure 5 shows the mean durations and F0 measurements of the target syllable followed by different following tones. Duration of the high-level tone does not significantly vary according to following tone, but F0 does. The effect of following tone on F0 of the high-level tone is significant at the offset of the F0 contour, [ $F(3, 32) = 8.57, p < 0.01$ ] and marginally significant at the mid-point of F0 contour, [ $F(3, 32) = 4.14, p < 0.05$ ]. As in the first pilot study, there was a dissimilation-like effect: the F0 values were higher when followed by syllables with low onsets (mid-falling or rising tone) than by syllables with mid or high onsets (high-level tone, mid-level tone or high-falling tone).

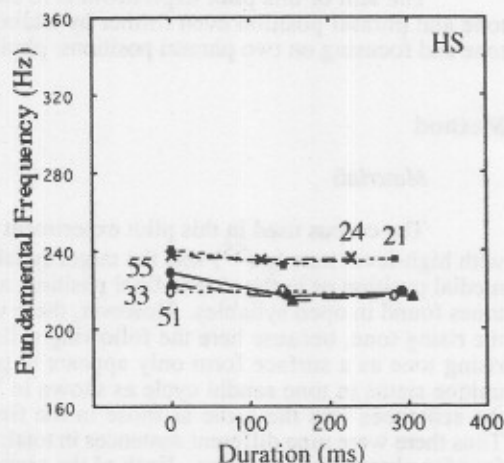


Fig. 5. Mean durations and F0s of the high target tone (55) followed by different tones and pooled over prosodic position.

## Main Study

The main experiment then was designed to investigate the issues raised in the pilot studies further by analyzing the productions of four subjects including all five Taiwanese tones occurring in non-checked syllables as targets in four prosodic positions and three tonal contexts.

## Method

### Materials

The corpus is shown in Appendix 3. Five target tones were investigated: [kaw<sup>55</sup>], [kaw<sup>33</sup>], [kaw<sup>21</sup>], [kaw<sup>51</sup>] and [kaw<sup>24</sup>]. Each of the target syllables was followed by high-level tone (55), mid-level tone (33) or mid-falling tone (21), and was placed in four different prosodic positions: phrase-initial, phrase-medial, phrase-final and utterance-final positions except that [kaw<sup>24</sup>] could occur only in

final position see Table 2. The utterances used in this study were five-syllable sentences and phrases and had the same prosodic structures used in the pilot studies.

There were forty-four utterances in total used in this study. Each of the sentences and phrases were read ten times. There were forty-four filler sentences mixed into the test sentences and phrases.

### *Subjects*

Four native Taiwanese speakers participated the test: two male and two female. They also speak Mandarin and English. None of them was the subject in the pilot experiments.

### *Recording and Acoustic Measurements*

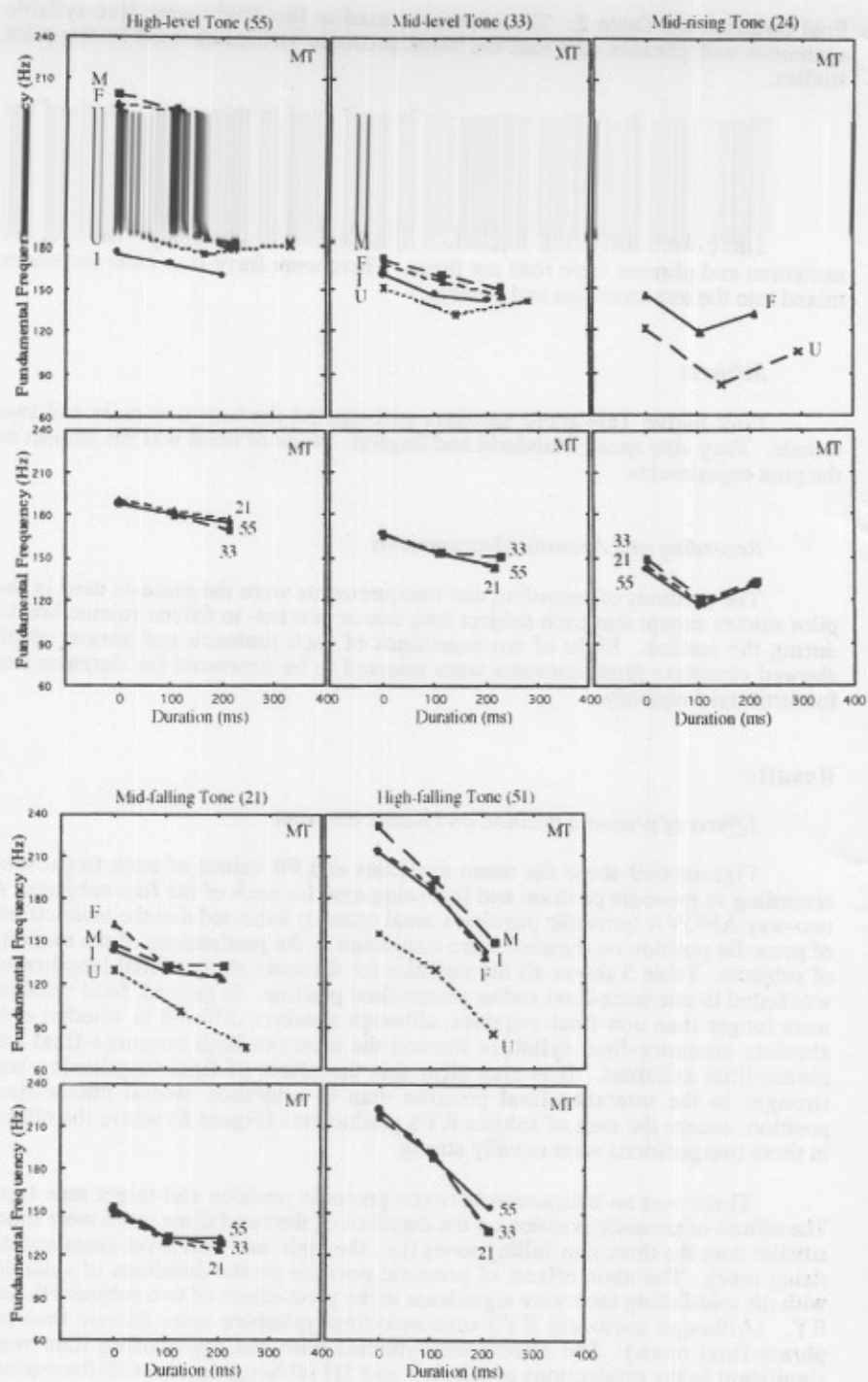
The methods of recording and measurements were the same as used in the pilot studies except that each subject took one or two ten- to fifteen-minute breaks during the session. Eight of ten repetitions of each sentence and phrase which showed complete pitch contours were selected to be measured for duration and fundamental frequency.

## **Results**

### *Effects of prosodic position on syllable duration*

Figures 6-9 show the mean durations and F0 values of each target tone according to prosodic position and following tone for each of the four subjects. A two-way ANOVA (prosodic position x tonal context) indicated that the main effects of prosodic position on duration were significant in the productions of the majority of subjects. Table 5 shows all the statistics for the main study. Final-lengthening was found in utterance-final and/or phrase-final position. In general, final syllables were longer than non-final syllables, although speakers differed in whether only absolute utterance-final syllables showed the effect or both utterance-final and phrase-final syllables. It is also clear that the effect of final-lengthening was stronger in the utterance-final position than in utterance medial phrase-final position, except the case of subject RY's productions (Figure 8) where the effects in these two positions were equally strong.

There was an interaction between prosodic position and target tone type. The effects of prosodic position on the durations of the two falling tones were much smaller than the three non-falling tones (i.e., the high- and mid-level tones and the rising tone). The main effects of prosodic position on the durations of syllables with the mid-falling tone were significant in the productions of two subjects HL and RY. (Although curiously RY's utterance-final syllables were shorter than his phrase-final ones.) The effects for syllables with the high-falling tone were significant in the productions of HL, RY and SH (although again with the curious short utterance-final syllables for RY) and was marginally significant in the production of subject MT. Some final-lengthening effect in the utterance-final position the were shown in the productions of subjects SH (Figure 7) and HL



**Fig. 6.** Mean durations and F0s of each target tone followed by different tones in different prosodic positions. Subject MT. I = phrase-initial, M = phrase-medial, F = phrase-final, U = utterance-final.

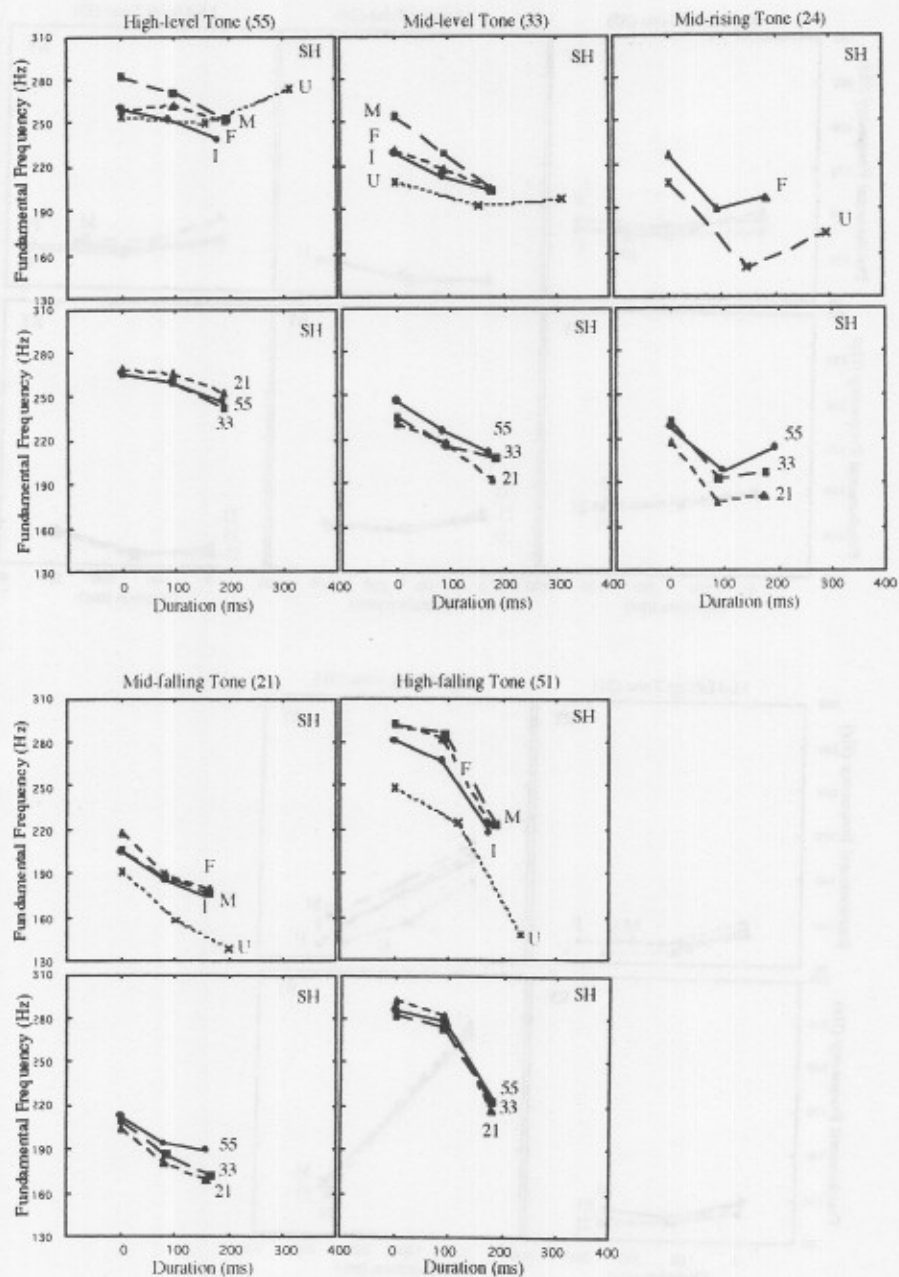


Fig. 7. Mean durations and F0s of each target tone followed by different tones in different prosodic positions. Subject SH. I = phrase-initial, M = phrase-medial, F = phrase-final, U = utterance-final.

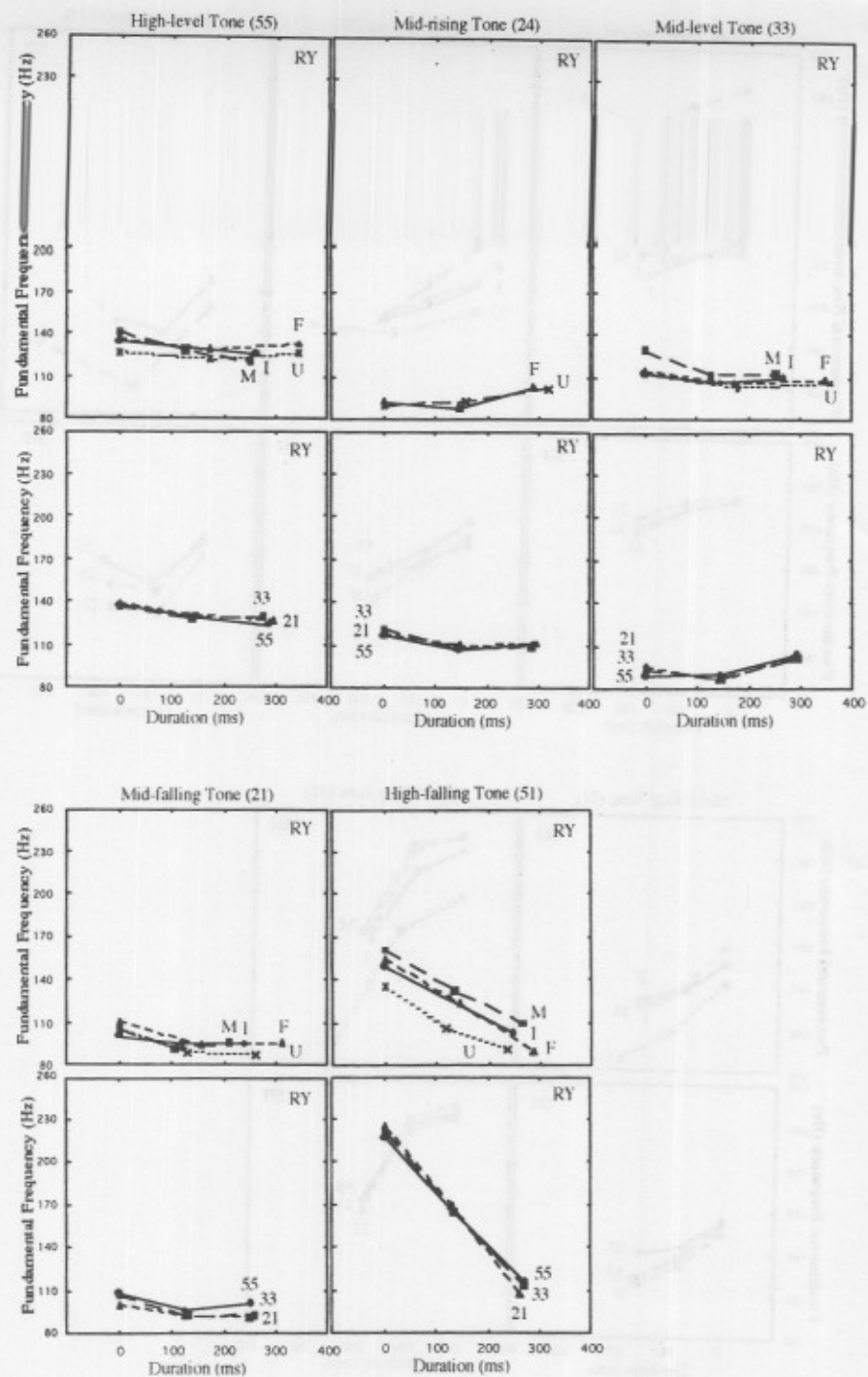


Fig. 8. Mean durations and F0s of each target tone followed by different tones in different prosodic positions. Subject RY. I = phrase-initial, M = phrase-medial, F = phrase-final, U = utterance-final.



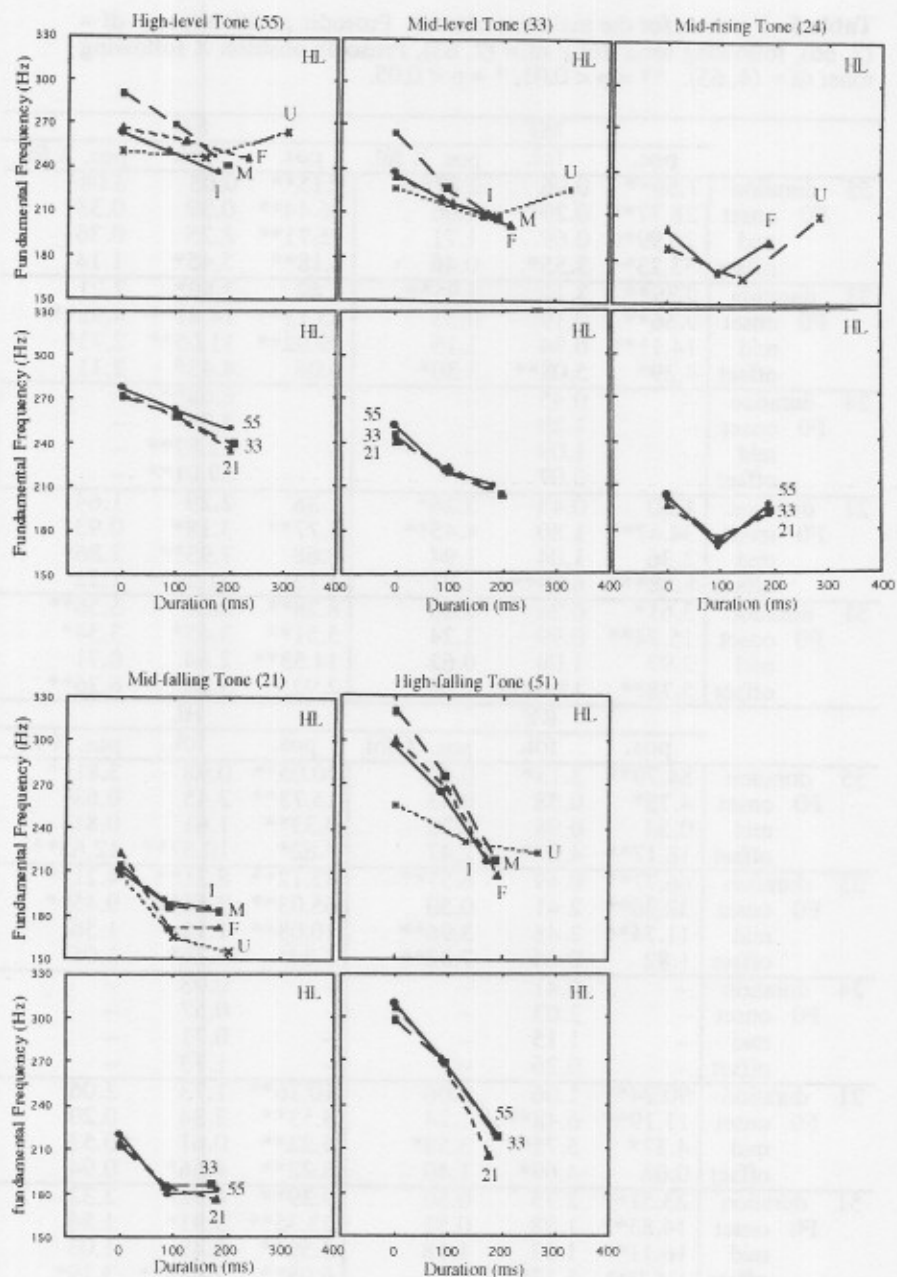


Fig. 9. Mean durations and F0s of each target tone followed by different tones in different prosodic positions. Subject HL. I = phrase-initial, M = phrase-medial, F = phrase-final, U = utterance-final.

**Table 5.** F values for the main experiment. Prosodic position (pos.): df = (2, 66), following tone (fol.): df = (2, 63), Prosodic position X following tone: df = (4, 63). \*\* = p < 0.01, \* = p < 0.05.

		MI			SH		
		pos.	fol.	pos. X fol.	pos.	fol.	pos. X fol.
55	duration	7.59**	0.06	2.87*	9.15**	0.05	3.08*
	F0 onset	28.77**	0.29	1.06	26.44**	0.39	0.33
	mid	26.99**	0.69	1.71	15.71**	2.25	0.36
	offset	33.23**	3.55*	0.46	8.18**	3.45*	1.14
33	duration	9.39**	3.12	3.95**	1.42	3.89*	2.21
	F0 onset	9.56**	0.19	0.33	53.13**	14.41**	4.92**
	mid	14.11**	0.34	2.15	29.02**	11.05**	2.73*
	offset	4.39*	5.08**	3.30*	0.06	4.43*	2.11
24	duration	—	0.45	—	—	6.68**	—
	F0 onset	—	1.29	—	—	4.34*	—
	mid	—	1.09	—	—	12.87**	—
	offset	—	0.07	—	—	19.91**	—
21	duration	1.80	0.43	3.26*	1.36	2.29	1.64
	F0 onset	34.47**	1.80	4.45**	8.77**	3.18*	0.93
	mid	2.36	1.04	1.94	0.68	7.93**	2.86*
	offset	8.88**	6.29**	1.91	1.12	21.29**	1.22
51	duration	3.61*	0.28	0.08	8.28**	0.29	5.56**
	F0 onset	15.74**	0.99	2.24	5.51**	3.65*	3.34*
	mid	2.97	1.09	0.62	14.53**	2.64	0.71
	offset	5.78**	15.41**	2.06	2.92	3.24*	6.26**
		RY			HL		
		pos.	fol.	pos. X fol.	pos.	fol.	pos. X fol.
55	duration	84.79**	3.15*	0.20	40.05**	0.98	3.81**
	F0 onset	4.75*	0.38	0.43	15.73**	2.45	0.63
	mid	0.54	0.98	0.70	9.33**	1.61	0.81
	offset	18.17**	4.07*	2.47	3.62*	10.57**	12.65**
33	duration	66.77**	0.49	6.57**	42.12**	8.21**	4.21**
	F0 onset	32.36**	2.41	0.50	65.03**	7.66**	9.45**
	mid	11.74**	2.48	3.96**	10.03**	1.12	1.56
	offset	1.82	0.64	7.32**	3.24*	3.20*	1.05
24	duration	—	0.41	—	—	0.03	—
	F0 onset	—	2.03	—	—	0.57	—
	mid	—	1.15	—	—	0.71	—
	offset	—	0.26	—	—	1.73	—
21	duration	90.24**	1.06	1.06	10.36**	1.73	2.06
	F0 onset	11.19**	6.48**	1.24	8.53**	2.34	0.20
	mid	4.37*	5.75**	3.59*	5.22**	0.61	0.52
	offset	0.08	4.69*	1.40	5.22**	4.66*	0.94
51	duration	23.51**	2.75	0.50	9.29**	4.92*	2.32
	F0 onset	14.83**	1.98	0.57	23.35**	3.91*	1.86
	mid	16.11**	1.37	1.18	5.59**	0.49	1.03
	offset	136.33**	5.17**	1.43	6.08**	10.88**	3.19*

(Figure 9). Final-lengthening of the falling tone in the phrase-final position only occurred in the productions of subject RY (Figure 8) and subject HL (Figure 9).

#### *Effects of prosodic position on F0*

Analysis of variance showed that the effect of prosodic position on F0s of target tones was significant for both the mid-points and offsets of F0 contours with only one exception—the offset of the mid-level tone produced by subject SH (Figure 7). Final-lowering of F0 occurred in utterance-final position and very weakly in phrase-final position. The effect on contour tones was stronger than that on level tones. The mid-points and offsets of the contour tones in the utterance-final position were lower than those in the other positions. In fact, the changes of mid-points and offsets of F0 contours also caused the pitch range of the contour tones to shift down. Figures 6-9 indicate that pitch ranges of contour tones in the utterance-final position, especially in the productions of subjects MT (Figure 6) and SH (Figure 7), were mostly lower than in other positions. Final-lowering in the phrase-final position was only found in the offsets of the high-falling tone in the production of subjects RY (Figure 8) and those of both falling tones in the productions of subject HL (Figure 9). The offsets of the high-falling tone in the phrase-medial position produced by subjects MT and RY and that of the mid-falling tone produced by MT were higher than those in the other positions. For the high-level tone and the mid-level tone, final-lowering was found in mid-point values in the utterance-final position. The offsets of the high-level tone and the mid-level tone in the utterance-final position were not lower than in other positions due to the change of tone shape except the mid-level tone shown in Figure 8 which was slightly lower than in phrase-medial position.

The level tones showed some change of tone shapes in the utterance-final or phrase-final positions. Tone shapes of contour tones: mid-falling, high-falling and rising tones were more constant. The only change found was in the high-falling tone in the utterance-final position produced by subject HL.

#### *Effects of tonal context on syllable duration*

The effect of following tone on durations of target tones was much smaller than that of prosodic position. Generally, syllable duration was constant regardless of the tonal context except for small differences in the productions of subject SH between mid-level tone vs. rising tone context (Figure 7), subject RY for high-level tone vs. high-falling tone contexts (Figure 8) and, subject HL for mid-level tone vs. high-falling tone context (Figure 9).

#### *Effects of tonal context on F0*

F0 values of target tones varied significantly according to following tones, although the effect was smaller than that of prosodic position. The main effect of following tone on F0 offsets was stronger than on mid-point F0s. Tonal assimilation occurred between the falling tones (51 & 21) and the following tones in the productions of all the subjects. The offset of the target tone was higher when followed by high-onset tone (55) than when followed by low-onset tone (21). This assimilatory effect was also found in the level tones (Figures 7 & 9) and in the rising tone (Figure 7). On other hand, the mid-point values were mostly constant except some small variation found in the productions of subjects SH and RY. A dissimilation-like effect of following tone was shown in the two level tones. The

F0 offset of the high-level tone or the mid-level tone in the context of the mid-level tone was higher than those in the context of the high-level tone (Figures 6 & 8). The F0 offset of the high-level tone in the context of the mid-falling tone was higher than that in the context of the mid-level tone (Figures 6 & 7).

## Discussion

As found in previous studies for other languages, durations of syllables varied according to prosodic position. Pre-pausal lengthening was the feature shared by all subjects. Duration of syllables was longer in the utterance-final position than in any other position probably due to the fact that speech speed was gradually reduced toward the post-utterance pause to signify the termination of the utterance. Consequently, syllables were lengthened utterance-finally. Klatt's study of connected discourse indicated that the duration of a vowel at the end of an utterance was not greater than at the end of a phrase. Nevertheless, he also predicted that when sentences are produced separately, the result may be different. Phrase-final lengthening was also present in the current study; however, the effect was not as strong as pre-pausal lengthening. It was consistently found only in the production of two subjects (HS and RY).

Lindblom and Rapp's (1973) prediction that segments in phrase-medial position would be shorter than those in phrase-initial and phrase-final position only partially agreed with the result found in the present study. Syllables in the phrase-medial position were shorter than in phrase-final position in the productions of most of the subjects. They were generally as long as syllables in phrase-initial positions with some variation. It seemed that the difference among syllables in these three phrasal positions could be appropriately interpreted by the substantial lengthening of syllables in the phrase-final position to signify the separation of phrases, although in fast speech the difference of the syllable duration in different phrasal positions might be accounted for by the speed of the utterance. Syllable duration was not influenced by the following tone except in very few cases showing small effects of following tone which were much weaker than the effect of prosodic position. That is, tonal context which varied the fundamental frequency did not temporally influence the duration of syllables.

The effect of prosodic position on duration of syllables was not equally great for all target tones. The durations of falling tones, especially the high-falling tone, were more constant than those of the two level tones and the rising tone. The temporal constancy of syllables with a falling tone indicated that the rate of F0 fall was probably important for the listeners to identify the quality of the tone. A study of perception of Taiwanese tones (Lin & Repp, 1989) showed that in addition to F0, duration of syllables also contributed to the distinction of falling and non-falling tones. Syllables with falling tone were found to be shorter than those with other tones in Taiwanese (Lin, 1988) as well as in Mandarin (Ho, 1976).

Mid-point and offset F0 values and the overall pitch range of tone were affected by both the prosodic position and the following tone. Final-lowering occurred in the utterance-final position as well as in the phrase-final position. However, the strongest effect was found in the utterance-final position. It is fairly possible that, in addition to the reduction of speech speed, the lowering of pitch also contributed to the cueing of utterance termination. The influence of phrasal position seemed to be stronger at the end of the syllable than at the middle of the syllable. The lowering of the offsets was greater than the mid-points. The


lowering of both mid-point values and offset values led to the lowering of the overall pitch range. The pitch ranges of tonal contours were shifted down in the utterance-final position regardless of the quality of the target tone. However, the lowering was greater for contour tones than level tones. It was probably due to the closer specification of register height for the level tones than for the contour tones. If the lowering of the F0 of the high-level tone (55) was too much, then the high-level tone was likely to be indistinguishable from the mid-level tone. Similarly, if the F0 of the mid-level tone was over-lowered, it would be difficult to distinguish it from the mid-falling tone which was sometimes analyzed as low tone (22 or 11). The lowering of the F0 height of falling or rising tones, on the contrary, was less likely to create this problem.

Phrase-final lowering was found in the production of some subjects, but it showed more variance and less effect than the lowering of F0 in the utterance-final position. The difference between the two final positions in the effect of final lowering was probably due to the fact that the F0 height in the phrase-final (non-utterance-final) could not be too low, since there needed to be some indication that the utterance will continue. Another reason was that given the short interval between prosodic phrases, the process of falling needed to be finished in a short period of time. Consequently, the F0 height could not be lowered as much as in the utterance-final position. F0 values of syllables in phrase-medial position and in phrase-initial position were higher than in final positions. The difference between the phrase-medial position and phrase-initial position found in very few cases was much less than the difference between non-final positions and final positions and showed much more variance. In these cases, F0 of tones in general was slightly higher in phrase-medial position than in phrase-initial position. This was different from what has been shown for Mandarin (Shih, 1988) where F0 is higher in utterance-initial position than in other positions in the utterance. This difference was probably caused by the slight difference in the prosodic position of the target tone. The target tone in the phrase-initial position used in the present study was not utterance-initial.

Unlike pitch range, tonal shape only showed very small and limited differences. The shape of some tonal contours in the utterance-final position were changed. All of them had the F0 which were raised at the offsets of the tonal contours. This phenomenon was found in the productions of four out of five subjects. The subject who participated the pilot studies did not show this change. The difference was possibly caused by the fact that the utterances used for the utterance-final case were all noun phrases in the pilot study and were imperative sentences (in some cases) in the main study. The subject whose production did not show change of tonal shape read the noun phrases as if naming objects. In the main study, subjects who raised their pitch at the end of the utterance (high boundary tone) read the imperative sentences in the way that meant to persuade someone to do something, instead of ordering someone to do something. (Higher pitch—or the lack of final lowering—sounds less like a declarative direct order.)

F0s of syllables were also affected by following tone, but the degree of the influence of the following tone was smaller than that of the prosodic position. As was seen in the effects of prosodic position, the offsets of the tonal contours showed a larger change than mid-points of F0s. Also, non-falling tones were slightly more stable across tonal contexts than were falling tones. The rising tone was the most stable tone. An assimilation effect was found between falling tones and tones following them. The offset values of falling tones change according to the height of the following tone. When the following tone was a high-onset tone,

the offset of the target tone was higher. It was lower when followed by a low-onset tone. A similar assimilatory effect was also found in anticipatory and perseveratory tonal coarticulation of Mandarin (Shih, 1988; Shen, 1990): tone 3 (214) starts higher when following a high tonal target corresponding to a high



offset than a low tonal target corresponding to a low offset. However, Shen found this effect to be the general trend for tonal coarticulation in Mandarin. In contrast, the results of the current study on Taiwanese showed also a dissimilation-like effect between level tones and tones following them. The F0 of a level tone was higher when followed by a low-onset tone, but was lower when followed by a high-onset tone. A similar effect was found in Shih's (1988) Mandarin study: tone 2 (245) ended higher when the following tonal target was low, but lower when the following tonal target was high. Shih accounted for this phenomenon in Mandarin as an artifact of target deletion and not of target value per se. She posited a rule that deletes the final high tonal target of tone 2 when the following tonal target is high so that the measured value is in the transition from the preceding mid to a high in the following syllable. By contrast, the anticipatory dissimilation of the offset of target tone in the present study occurred in level tones and cannot be explained by a shift in timing of tonal target. It seems to be a contrast effect occurring between tones with different onset and offset values either to enhance perceptual cues of the following tone or to preserve the contrast between the high-level tone and the high-falling tone.

Pitch range and tone shape of tones were less affected by the following tone than by prosodic position. The change of overall pitch level caused by the F0 height of the following tone was only found in the high-level tone produced by subject HS. This minor change of pitch range of tone by tonal context was similar to that found in Lin's Taiwanese study (1988), but differed from the pattern found in Mandarin which showed change of pitch range of tones depending on the tonal context except for the high-falling tone only affected at the offset. In the production of HS, not only the offsets of tones but also the mid-point F0s were affected by the following tone. As found in Mandarin (Shen, 1990), tone shape was generally constant regardless of different following tones except a small change found in the production of subject HS.

## Conclusion

Taiwanese tones are acoustically influenced by prosodic position and to a certain extent by tonal context. Prosodic position has strong effects on both the F0 and duration of tones. Final-lengthening and final-lowering of F0 were found in the utterance-final and to a somewhat lesser extent in phrase-final positions. Pitch range was substantially affected by prosodic position, but in general, the tone shape was not changed. On the other hand, anticipatory tonal coarticulation affects the F0 offsets of tones without affecting syllable duration. A dissimilation-like effect occurs between level tones and the following tone. The F0 offset of a level tone was higher when the onset of the following tone was low than when it was high. Assimilation occurs between contour tones and the following tone. Some assimilatory effects were also found in level tones. The F0 offset of a contour tone was higher when followed by a high-onset tone than when followed by a low-onset tone. Tone shape of Taiwanese tones is not affected by tonal context. Pitch range in general is not affected by tonal context; however, some subject-dependent variation was found.

Given the effects of prosodic position and tonal context on duration and the F0 contours of tones, it would be interesting to see how duration and F0 contribute to listener identification of the prosodic position and the quality of the following tone. What are the relative contributions of the temporal information and the F0 cuing the termination or continuation of an utterance? Can the quality of a tone be predicted from the preceding tone?

## References

- Chao, Y. R. (1930) A system of tone letters. *Le Maitre Phonétique*, 30, 24-27.
- Chen, M. Y. (1987) The syntax of Xiamen tone sandhi. *Phonology Yearbook*, 4, 109-149.
- Cheng, R. (1968) Tone sandhi in Taiwanese. *Linguistics*, 41, 19-42.
- Cheng, R. (1973) Some notes on tone sandhi in Taiwanese. *Linguistics*, 100, 5-25.
- Entropic Research Laboratory, INC. (1993) *Waves+ 5.0*. AT & T Bell Laboratories. Washington, D. C..
- Fonagy, I. & Magdics, K. (1960) Speed of utterance in phrases of different lengths. *Language and Speech*, 3, 179-192.
- Ho, A. T. (1976) The acoustic variation of Mandarin tones. *Phonetica*, 33, 353-367.
- Hsiao, Y. E. (1990) The Bermuda triangle of syntax, rhythm and tone. *ESCOL '90*, 112-123.
- Klatt, D. (1975) Vowel lengthening is determined in a connected discourse. *Journal of Phonetics*, 3, 129-140.
- Liberman, M. & Pierrehumbert, J. (1984) Intonational invariance under changes in pitch range and length. In *Language Sound Structure*, (M. Aronoff & R. T. Oehrle, editors), pp. 157-233. Cambridge, MA; MIT Press.
- Lin, H.-B. (1988) *Contextual Stability of Taiwanese Tones*. Doctoral dissertation, University of Connecticut.
- Lin, H.-B. & Repp, B. (1989) Cues to the perception of Taiwanese tones. *Language and Speech*, 32(1), 25-44.
- Lindblom, B. & Rapp, K. (1973) Some temporal regularities of spoken Swedish. *Papers from the Institute of Linguistics, University of Stockholm, Publication 21*.
- Shen, X. (1990) Tonal coarticulation in Mandarin. *Journal of Phonetics*, 18, 281-295.
- Shih, C. (1988) Tone and intonation in Mandarin. *Working Papers of the Cornell Phonetics Laboratory*, 3, 83-109.

Appendix 1—Corpus for first pilot study

[kɔ<sup>55</sup>] followed by high-level tone (55)

[t<sup>h</sup>en<sup>55</sup> po<sup>55</sup>][kɔ<sup>55</sup> tɔŋ<sup>55</sup> tiam<sup>21</sup>]

"TAN PAW" antique shop.

[twa<sup>21</sup> kɔ<sup>55</sup> tɔŋ<sup>55</sup>] [bo<sup>33</sup> lai<sup>24</sup>]

The great shareholder did not come.

[tam<sup>33</sup> hiŋ<sup>33</sup> kɔ<sup>55</sup>] [kau<sup>55</sup> kin<sup>55</sup>]

There are nine Chinese pounds of fresh mushrooms.

[tam<sup>33</sup> hiŋ<sup>33</sup> kɔ<sup>55</sup>]

fresh mushrooms

[kɔ<sup>55</sup>] followed by mid-level tone (33)

[te<sup>33</sup> kɔ<sup>55</sup> bi<sup>33</sup>] [tsin<sup>33</sup> taŋ<sup>33</sup>]

This has a strong smell of a kettle.

[tam<sup>33</sup> hiŋ<sup>33</sup> kɔ<sup>55</sup>] [tɿ<sup>33</sup> tɿ<sup>33</sup>]

Fresh mushrooms are slightly sweet.

[tam<sup>33</sup> hiŋ<sup>33</sup> kɔ<sup>55</sup>]

fresh mushrooms

[pɔ<sup>51</sup>] followed by high-level tone (55)

[si<sup>51</sup> tsi<sup>51</sup>] [pɔ<sup>51</sup> kau<sup>55</sup> tswa<sup>33</sup>]

At least nine rows are planted.

[t<sup>h</sup>sa<sup>33</sup> pɔ<sup>51</sup> pi<sup>55</sup>] [tan<sup>51</sup> tsi<sup>24</sup>]

Sewing clothes is a way to earn some money.

[tsen<sup>33</sup> t<sup>h</sup>sai<sup>51</sup> pɔ<sup>51</sup>] [kan<sup>55</sup> tan<sup>55</sup>]

It is easy to fry some dried radish.

[tsen<sup>33</sup> t<sup>h</sup>sai<sup>51</sup> pɔ<sup>51</sup>]

To fry some dried radish.

Appendix 2—Corpus for second pilot study

[pɔ<sup>55</sup>] in the phrase-medial position with five different following tone

[twa<sup>21</sup> pɔ<sup>55</sup> to<sup>55</sup>] [tsin<sup>33</sup> lai<sup>33</sup>]

The big valuable knives are very sharp.

[twa<sup>21</sup> pɔ<sup>55</sup> ten<sup>33</sup>] [p<sup>h</sup>aɪ<sup>55</sup> k<sup>h</sup>i<sup>51</sup>]

Large precious temples are hard to build.

[twa<sup>21</sup> pɔ<sup>55</sup> kiam<sup>21</sup>] [si<sup>51</sup> kwi<sup>21</sup>]

The big precious sword is most expensive.

[twa<sup>21</sup> pɔ<sup>55</sup> tia<sup>51</sup>] [kia<sup>55</sup> hwe<sup>55</sup>]

The big precious caldron is not fire-proof.

[twa<sup>21</sup> pɔ<sup>55</sup> k<sup>h</sup>im<sup>24</sup>] [tat<sup>21</sup> tsi<sup>24</sup>]

Big precious pianos are worth a lot.

[pɔ<sup>55</sup>] in the phrase-final position with four different following tone

[bo<sup>33</sup> laŋ<sup>33</sup> pɔ<sup>55</sup>] [k<sup>h</sup>ɔ<sup>55</sup> lien<sup>24</sup>]

It is depressing to be praised by nobody.



[ŋ <sup>51</sup> laŋ <sup>33</sup> po <sup>55</sup> ] [tsin <sup>33</sup> t <sup>h</sup> sa <sup>m</sup> <sup>51</sup> ]	It is awful to expect praise from people.
[u <sup>21</sup> laŋ <sup>33</sup> po <sup>55</sup> ] [si <sup>521</sup> ho <sup>51</sup> ]	It is good to be praised by some people.
[ai <sup>51</sup> laŋ <sup>33</sup> po <sup>33</sup> ] [kio <sup>51</sup> kak <sup>21</sup> ]	It is hopeless to be always eager for praise.

((σ σ \*po<sup>33</sup>) [σ<sup>24</sup> σ] missing since 24 occurs only in final position--see Table 2)

### Appendix 3—Corpus for main experiment

[kau<sup>55</sup>] in the phrase-initial position

[lɔŋ <sup>55</sup> tsɔŋ <sup>51</sup> ] [kau <sup>55</sup> taŋ <sup>55</sup> puā <sup>21</sup> ]	In total, there are nine and a half barrels.
[lɔŋ <sup>55</sup> tsɔŋ <sup>51</sup> ] [kau <sup>55</sup> kin <sup>33</sup> puā <sup>21</sup> ]	In total, there are nine and a half Chinese pounds.
[lɔŋ <sup>55</sup> tsɔŋ <sup>51</sup> ] [kau <sup>55</sup> pɔ <sup>21</sup> puā <sup>21</sup> ]	In total, there are nine and a half pounds.

[kau<sup>55</sup>] in the phrase-medial position

[be <sup>55</sup> kau <sup>55</sup> kin <sup>55</sup> ] [bo <sup>33</sup> kau <sup>51</sup> ]	It is not enough to just buy nine Chinese pounds.
[be <sup>55</sup> kau <sup>55</sup> pɔ <sup>33</sup> ] [bo <sup>33</sup> kau <sup>51</sup> ]	It is not enough to just buy nine pounds.
[be <sup>55</sup> kau <sup>55</sup> te <sup>21</sup> ] [bo <sup>33</sup> kau <sup>51</sup> ]	It is not enough to just buy nine pieces.

[kau<sup>55</sup>] in the phrase-final position

[si <sup>33</sup> ban <sup>21</sup> kau <sup>55</sup> ] [kiam <sup>55</sup> ke <sup>21</sup> ]	The price will be reduced if it is turned in late.
[si <sup>33</sup> ban <sup>21</sup> kau <sup>55</sup> ] [ka <sup>33</sup> kɔ <sup>24</sup> ]	The payment will be increased if it is turned in late.
[si <sup>33</sup> ban <sup>21</sup> kau <sup>55</sup> ] [po <sup>21</sup> li <sup>21</sup> ]	The interest will be low if it is turned in late.

[kau<sup>55</sup>] in the utterance-final position

[m <sup>21</sup> t <sup>h</sup> aŋ <sup>33</sup> si <sup>33</sup> ban <sup>21</sup> kau <sup>55</sup> ]	Do not turn it in late.
---	-------------------------

[kau<sup>33</sup>] in the phrase-initial position

[tsi <sup>21</sup> kai <sup>51</sup> ] [kau <sup>33</sup> kau <sup>55</sup> pa <sup>21</sup> ]	Nine hundred dollars is paid each time.
[tsi <sup>21</sup> kai <sup>51</sup> ] [kau <sup>33</sup> te <sup>33</sup> puā <sup>21</sup> ]	Nine and a half bags are submitted each time.

[tsi <sup>21</sup> kai <sup>51</sup> ] [kau <sup>33</sup> pɔ <sup>55</sup> pu <sup>21</sup> ]	Nine and a half pounds are submitted eachtime.
[kau <sup>33</sup> ] in the phrase-medial position	
[kh <sup>51</sup> kau <sup>33</sup> kh <sup>55</sup> u <sup>55</sup> ] [sam <sup>51</sup> pɔ <sup>33</sup> ]	(He goes) to the suburbs to take a walk.
[kh <sup>51</sup> kau <sup>33</sup> kwu <sup>33</sup> ] [sam <sup>51</sup> pɔ <sup>33</sup> ]	(He goes) to the outskirts of the city to take a walk.
[bo <sup>33</sup> kau <sup>33</sup> tai <sup>21</sup> ] [to <sup>33</sup> tsau <sup>51</sup> ]	(He) left without leaving a message.
[kau <sup>33</sup> ] in the phrase-final position	
[tso <sup>51</sup> si <sup>33</sup> kau <sup>33</sup> ] [p <sup>h</sup> ai <sup>55</sup> pau <sup>55</sup> ]	If (the skin of a dumpling) is made too thick, it will be hard to seal.
[tso <sup>51</sup> si <sup>33</sup> kau <sup>33</sup> ] [kia <sup>55</sup> ke <sup>55</sup> ]	If (the coat) is made too bulky, it will be inconvenient.
[tso <sup>51</sup> si <sup>33</sup> kau <sup>33</sup> ] [p <sup>h</sup> ai <sup>55</sup> pau <sup>55</sup> ]	If (the skin of a dumpling) is made too thick, it will weigh a lot.
[kau <sup>33</sup> ] in the utterance-final position	
[m <sup>21</sup> th <sup>aŋ</sup> <sup>33</sup> tso <sup>51</sup> si <sup>33</sup> kau <sup>33</sup> ]	Do not make it too thick.
[kau <sup>21</sup> ] in the phrase-initial position	
[nai <sup>21</sup> iɔŋ <sup>33</sup> ] [kau <sup>21</sup> ti <sup>55</sup> kwa <sup>21</sup> ]	durable thick lids of cookers
[nai <sup>21</sup> iɔŋ <sup>33</sup> ] [kau <sup>21</sup> pɔ <sup>33</sup> le <sup>24</sup> ]	durable thick glass
[nai <sup>21</sup> iɔŋ <sup>33</sup> ] [kau <sup>21</sup> te <sup>21</sup> pan <sup>51</sup> ]	durable thick floor
[kau <sup>21</sup> ] in the phrase-medial position	
[i <sup>521</sup> kau <sup>21</sup> paŋ <sup>55</sup> ] [lai <sup>33</sup> kh <sup>am</sup> <sup>21</sup> ]	(We should) cover it with a thick piece of wood.
[i <sup>521</sup> kau <sup>21</sup> p <sup>h</sup> we <sup>33</sup> ] [lai <sup>33</sup> kh <sup>am</sup> <sup>21</sup> ]	(We should) cover it with a thick comforter.
[i <sup>521</sup> kau <sup>21</sup> pɔ <sup>21</sup> ] [lai <sup>33</sup> kh <sup>am</sup> <sup>21</sup> ]	(We should) cover it with a thick cloth.
[kau <sup>21</sup> ] in the phrase-final position	
[be <sup>21</sup> bo <sup>33</sup> kau <sup>21</sup> ] [kiam <sup>55</sup> th <sup>an</sup> <sup>21</sup> ]	Less money was earned due to fewer materials being on sale.
[tsia <sup>21</sup> bo <sup>33</sup> kau <sup>21</sup> ] [tiŋ <sup>33</sup> kio <sup>21</sup> ]	If it is not enough to eat, (you should) order more.

[ten <sup>33</sup> bo <sup>33</sup> kau <sup>21</sup> ] [ten <sup>33</sup> tsi <sup>51</sup> ]	Fewer fish were caught due to the lack of electricity.
[kau <sup>21</sup> ] in the utterance-final position	
[tʰsin <sup>33</sup> tʰsi <sup>51</sup> tsi <sup>21</sup> bo <sup>33</sup> kau <sup>21</sup> ]	It does not seem to be enough to eat.
[kau <sup>51</sup> ] in the phrase-initial position	
[tsi <sup>21</sup> tʰiau <sup>21</sup> ] [kau <sup>51</sup> ti <sup>55</sup> kuan <sup>24</sup> ]	(He) jumped right to the top.
[tsi <sup>21</sup> tʰiau <sup>21</sup> ] [kau <sup>51</sup> tu <sup>33</sup> ti <sup>51</sup> ]	(He) jumped right to the top of the closet.
[tsi <sup>21</sup> tʰiau <sup>21</sup> ] [kau <sup>51</sup> pʰ we <sup>221</sup> ti <sup>51</sup> ]	(He) jumped right to the top of the comforter.
[kau <sup>51</sup> ] in the phrase-medial position	
[bo <sup>33</sup> kau <sup>51</sup> tɿ <sup>55</sup> ] [pʰ ai <sup>55</sup> tsi <sup>21</sup> ]	If it is not sweet enough, it will not be delicious.
[bo <sup>33</sup> kau <sup>51</sup> ti <sup>33</sup> ] [pʰ ai <sup>55</sup> tsi <sup>21</sup> ]	If it is not firm enough, it will not be tasteful.
[bo <sup>33</sup> kau <sup>51</sup> kʰ wi <sup>21</sup> ] [pʰ ai <sup>55</sup> tsi <sup>21</sup> ]	If (he) is not satisfied, (he) will not give up.
[kau <sup>51</sup> ] in the phrase-final position	
[gi <sup>21</sup> tsap <sup>21</sup> kau <sup>51</sup> ] [tam <sup>55</sup> gan <sup>51</sup> ]	The eyes (of the statue) will be painted on the 29th.
[gi <sup>33</sup> tsap <sup>21</sup> kau <sup>51</sup> ] [kʰ wi <sup>33</sup> tsi <sup>51</sup> ]	The result of the lottery will be known on the 29th.
[gi <sup>33</sup> tsap <sup>21</sup> kau <sup>51</sup> ] [ta <sup>21</sup> ka <sup>55</sup> ]	The construction will begin on the 29th.
[kau <sup>51</sup> ] in the utterance-final position	
[læ <sup>21</sup> pa <sup>21</sup> gi <sup>33</sup> tsap <sup>21</sup> kau <sup>51</sup> ]	Two hundred and twenty-nine.
(kau <sup>24</sup> cannot occur in phrase initial or medial position see Table 2)	
[kau <sup>24</sup> ] in the phrase-final position	
[hit <sup>53</sup> tsi <sup>51</sup> kau <sup>24</sup> ] [kau <sup>55</sup> kwai <sup>21</sup> ]	That monkey is nasty.
[hit <sup>53</sup> tsi <sup>51</sup> kau <sup>24</sup> ] [tʰ au <sup>33</sup> tsi <sup>33</sup> ]	That monkey ate without permission.
[hit <sup>53</sup> tsi <sup>51</sup> kau <sup>24</sup> ] [pʰ o <sup>33</sup> ki <sup>51</sup> ]	That monkey is carrying her baby.

[kau<sup>24</sup>] in the utterance-final position

[k<sup>h</sup>wā<sup>51</sup> tiō<sup>21</sup> hit<sup>53</sup> tsia<sup>51</sup> kau<sup>24</sup>] (I) saw that monkey.

## The influence of syntax on prosodic structure in Japanese\*

Jennifer J. Venditti

venditti@ling.ohio-state.edu

**Abstract:** This paper examines the relationship between the syntactic and prosodic structures of utterances which are structurally ambiguous. Two experiments were conducted involving ambiguous noun phrases (right- versus left-branching) and relative clause constructions containing an adjunct with ambiguous scope of modification. Results of careful examination of the F0 contours and downstepping patterns reveal that inter- and intra-speaker variability as well as depth of syntactic embedding are important factors in determining the prosodic phrasing.

### 1 Introduction

Many acoustic studies in recent years have been concerned with the role suprasegmental features play in helping cue the syntactic structure of a sentence (e.g. Cooper et al., 1978; Klatt, 1975; Lehiste, 1973; Lehiste et al., 1976; Streeter, 1978; Terken & Collier, 1992; among many others). There is a wealth of literature concerning the disambiguation of syntactic ambiguity, discussing acoustic features such as duration, amplitude and fundamental frequency (F0) which provide a crutch to cueing differing interpretations of an otherwise ambiguous string of segments. Klatt (1975) found in English that increased duration can mark the ends of major syntactic units. Streeter (1978) showed that both duration and pitch contour, as well as amplitude, serve as salient cues to determining syntactic boundaries in ambiguous algebraic expressions. Lehiste (1973, 1976) found that duration, and F0 to a certain extent, plays a major role in disambiguation of many syntactically ambiguous sentences.

Japanese is no different from the Indo-European languages in this respect: there has been much work done on ambiguity which shows that acoustic features can indeed cue the syntactic structure (e.g. Uyeno et al., 1979, 1980, 1981; Azuma & Tsukuma, 1990, 1991; Venditti & Yamashita, in preparation). A study done by Uyeno and her colleagues (1980) on relative clause constructions in Tokyo Japanese showed a distinct F0 contour for each of the interpretations of an ambiguous utterance. An example sentence from their corpus is given in (1).

- (1) Ototoi koronda otona-ga waratta  
day before yesterday fell adult-NOM laughed

A: 'The adult who fell the day before yesterday laughed.'

B: 'The adult who fell laughed the day before yesterday.'

---

\*The work reported in this paper was supported in part by an Ohio State University Center for Cognitive Science Interdisciplinary Summer Fellowship and a Title VI Foreign Language Area Studies Fellowship. I would like to thank Mary Beckman, Beth Hume, Stefanie Jannedy, Keith Johnson, Sun-Ah Jun, Andreas Kathol, Mineharu Nakayama, and Frederick Parkinson for their helpful discussions, and Azusa Morii for her native speaker talents and endless cooperation.

From a perception test using synthesized F0 contours containing no pauses, it was shown that when the F0 peak on the adverb *ototoi* 'the day before yesterday' (Peak 1) is a great deal higher than the peak on the verb *koronda* 'fell' (Peak 2), interpretation A (in which the adverb is modifying the verb of the relative clause) is preferred. On the other hand, manipulation such that Peak 2 is equal to or higher than Peak 1 will yield a tendency toward interpretation B, in which the initial adverb modifies the verb of the matrix clause. This result has been replicated by Azuma and Tsukuma (1990, 1991) for the Tokyo dialect, and for the Kinki dialect as well.

The studies mentioned above all assume that variations in the suprasegmental characteristics of the speech signal are direct manifestations of differences in syntactic structure. They all assume that the syntax influences the phonetic representation in a straightforward fashion, without any mediating levels of structure. However, work in the last decade on the relation between the phonetic representation and the surface syntax has suggested there is indeed an intermediate level: the prosodic structure (e.g. Selkirk, 1984, 1986; Nespor & Vogel, 1986; Kubozono, 1988; among many others). The proposal of such a structure was motivated by observations that the domains of various phonological rules are not isomorphic to the syntactic structure, but in fact only correspond to the syntax in an indirect manner. This has in turn led to a handful of syntax-prosody mapping algorithms, as will be briefly discussed in §5. It is generally accepted now that there is a prosodic structure which is made up of hierarchically organized levels of phrasing, each of which corresponds to the domains of phonological phenomenon like downstep application in English or Japanese, or postlexical gemination in Italian. Prosodic constituents at each level of the hierarchy are independently motivated according to which phonological rules apply within their domain.

The existence of an intermediate prosodic structure which is only indirectly related to the syntactic structure creates the need to reconsider the results of previous acoustic studies on disambiguation cited above. It is apparent that the syntactic structure in some way has an effect on the phonetic output of the utterance, but *how* it has this effect is not clear. Do the different interpretations of ambiguous sentences have distinct prosodic representations? If so, at what levels in the prosodic hierarchy are the two distinct? The present study attempts not only to show that there are indeed differences in the speech signals in ambiguous constructions, but also to examine whether these are different in terms of their prosodic representations, and if so, to determine which levels of the hierarchy are relevant for disambiguation in Japanese.

## 2 The Prosodic Hierarchy in Japanese

Before proceeding to description of the experiments and discussion of prosodic representations of ambiguous constructions, I will give a whirlwind tour of those parts of the prosodic hierarchy of Japanese which have, to a large extent, been agreed upon by many of those working with Japanese intonation (see Beckman & Pierrehumbert, 1986; Kubozono, 1988, 1989, 1992; Maekawa, 1991; McCawley, 1968; Poser, 1984; Pierrehumbert & Beckman, 1988; Selkirk and Tateishi, 1988, 1991; among others).

I will assume throughout this paper that the intonation pattern is intimately connected to the prosodic structure of an utterance, and that this structure is directly manifested by the surface realization of the fundamental frequency contour. Thus, by observing patterns in this contour, we are able to make claims about the prosodic organization of the utterance.

## 2.1 Lexically specified pitch accent

In Japanese, each lexical item is classified as accented or unaccented. An 'accented' word is one which has a bitonal H\*+L (henceforth HL) pitch accent associated to some specified mora. It is generally accepted that a word can have a maximum of one accent associated to it (see Poser (1990) for argument), and the location of the accent is determined at the lexical level. An 'unaccented' word, on the other hand, does not have this pitch accent associated to it, and is characterized mainly by tones associated with the accentual phrase level of the prosodic hierarchy (see §2.2). Consider the following minimal pair:

- (2) accented: ue'ru mono 'the ones that are starved'  
                  |  
                  HL  
unaccented: ueru mono 'something to plant'

(Henceforth I will transcribe an accented word with the diacritic (') after the mora to which the accent is associated.) Thus, the HL bitonal accent which is characteristic of accented words is assigned in the lexicon, while other tones which characterize a word, be it accented or unaccented, are assigned at the accentual phrase level of the intonational structure.

## 2.2 The accentual phrase: Phrasal and boundary tones

The accentual phrase is a level of the prosodic hierarchy which is essentially the same entity as the 'minor phrase' discussed by other scholars. This level of prosodic phrasing is defined for Japanese as "the domain of a postlexical rule deleting all accents after the first in an accentual phrase, and, more important, it is the domain of two delimitative peripheral tones, the phrasal H and the boundary L%." (Pierrehumbert & Beckman, 1988:26) An accentual phrase is most commonly thought to consist of a word plus its postposition or case marker. However, it is quite possible for a sequence of more than one word to combine to form one accentual phrase delimited by the phrasal H and L% boundary tone, and optionally at most one lexically specified accent. Figure 1a shows an accentual phrase consisting of a two-word sequence with an accented lexical item *ue'ru* 'to starve'. The fundamental frequency contour (the manifestation of these tones) is characterized by an initial L% boundary tone inserted in absolute utterance-initial position, a H phrasal tone followed by a steep fall associated with the HL pitch accent, and finally by the L% boundary tone. Figure 1b shows the contour of a single accentual phrase consisting of a sequence with an unaccented lexical item *ueru* 'to plant'. Here, the fundamental frequency starts off low (utterance initial L%), rises to the H phrasal tone and gradually tapers off toward the final L% boundary tone. This H phrasal tone will be associated to the second mora (if it is a short syllable) of a word.

The grouping of words into accentual phrases is a complicated phenomenon which is beyond the scope of this paper. It is worth mentioning, however, that accented words tend to resist grouping with other accented words into a single accentual phrase, while unaccented words tend to group together more readily. This fact will become relevant below when the intonation contours are examined. (For a discussion of factors governing accentual phrase formation, see Kori (1992) for Japanese and Jun (1993) for a related discussion of Korean.)

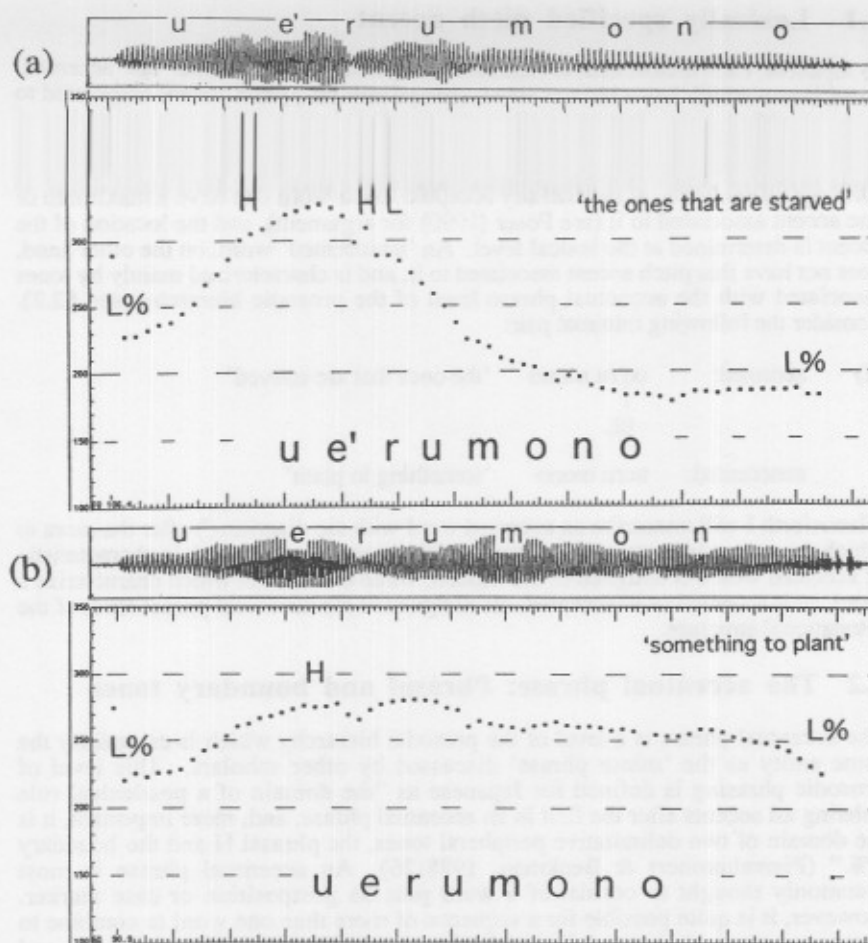


Figure 1. (a) An accentual phrase with an accented word *ue'ru mono* 'the ones that are starved'. (b) An accentual phrase with an unaccented word *ueru mono* 'something to plant'.

### 2.3 The major phrase: The domain of downstep

Much as in English and many African tone languages, pitch range in Japanese is manipulated by a process called downstep (or 'catathesis' as termed by Poser (1984) and Pierrehumbert & Beckman (1988)). In Japanese, downstep is the phonologically conditioned reduction in pitch range after the HL pitch accent. It has been suggested by Poser (1984) and confirmed by Pierrehumbert and Beckman (1988) and Kubozono (1988) that downstep applies iteratively within the bounds of a prosodically grouped set of accentual phrases. This larger prosodic constituent has been called the major phrase, and corresponds to Pierrehumbert and Beckman's 'intermediate phrase' which they compare to the analogous domain of downstep in English. Thus, in a string of accented accentual phrases that together form a major phrase, the pattern of the accentual phrase peaks will resemble a descending staircase. By definition, the contour of a string of unaccented accentual phrases



will not show the same pattern, since there are no accents to trigger the downstep process.

The level of the major phrase is extremely relevant to the present study, as the patterning of downstep between successive peaks will be examined in order to determine the proper prosodic phrasing of the utterances.

## 2.4 Contrasting theories of prosodic organization

In contrast to the widespread agreement on which levels of the prosodic hierarchy are relevant for Japanese, there is less uniformity of opinion about how the levels are organized with respect to each other. There are two main viewpoints: that held by Kubozono (1988, 1989, 1992) and that held by Beckman and Pierrehumbert (Beckman & Pierrehumbert, 1986; Pierrehumbert & Beckman, 1988).

Kubozono follows Ladd (1986) in assuming a recursive prosodic structure. That is, he explicitly rejects the Strict Layer Hypothesis (see Selkirk, 1984) by which prosodic units of type  $X^{n-1}$  are exhaustively grouped into units at the next higher level  $X^n$ , and instead proposes that prosodic constituents in Japanese are arranged in a binary branching hierarchical structure, which he gets by allowing embedding of prosodic constituents within constituents of the same type, shown in Figure 2a. This idea was originally proposed by Ladd for English to account for observed trends of 'declination within declination' — that is, downward trends of the fundamental frequency contour which seem to be embedded within larger downward trends. The idea was further developed by Ladd (e.g. 1988, 1990, 1993) into a metrical representation of prosodic structure and pitch register. In his model, the local prominences of pitch accents are represented as high and low terminal nodes of a metrical tree which can in turn be dominated by other high or low mother nodes to show prominence relationships among groups of accents. Such a representation attempts to account for relations among pitch registers by the combination of highs and lows and by the depth of embedding in the metrical tree. In such a model, there are necessarily a fixed number of pitch range values.

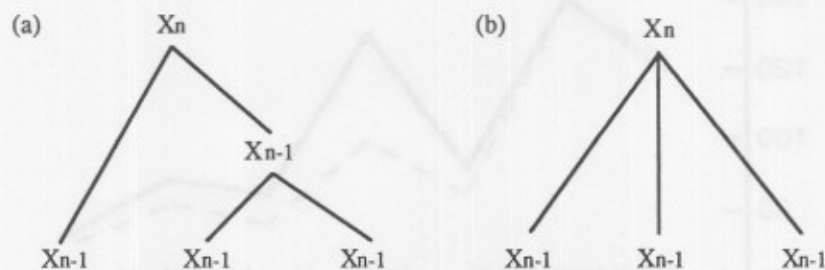


Figure 2. (a) Recursive versus (b) strictly layered hierarchical representations.

An alternate model of prosodic structure and pitch range relationships is described by Beckman and Pierrehumbert. They propose an  $n$ -ary branching hierarchical structure in which prosodic constituents group together according to the Strict Layer Hypothesis, as shown in Figure 2b. In this model, pitch range is a continuously variable phonetic specification chosen just once for any given major or 'intermediate' phrase, and at the beginning of a new major phrase, the pitch range is specified independently from that of the preceding phrase. Thus, the choice of pitch range for a given phrase is a paradigmatic one, reflecting the overall discourse structure or the pragmatic context. This forms a sharp contrast to the proposals of

Ladd and Kubozono, in which pitch range relationships are represented phonologically as a result of particular arrangements in the structure of their prosodic representation.

## 2.5 Differing accounts of ambiguous constructions

With these differences in mind, let us now turn to examples of syntactically ambiguous constructions.

Kubozono (1988, 1989, 1992) has studied extensively the relation between syntactic structure and fundamental frequency peak values for noun phrases with differing branching structures, such as the those given in (3). He also examined ambiguous strings such as in (4).

- (3) a. [[ao'yama-ni a'ru] daigaku] 'a university in Aoyama'  
           Aoyama-in exist university  
       b. [ao'yama-no [a'ru daigaku]] 'a certain university in Aoyama'  
           Aoyama-GEN certain university
- (4) a. [[o'okina nooen-no] o'onaa] 'an owner of a big farm'  
           big farm-GEN owner  
       b. [o'okina [nooen-no o'onaa]] 'a tall farm-owner'  
           big farm-GEN owner

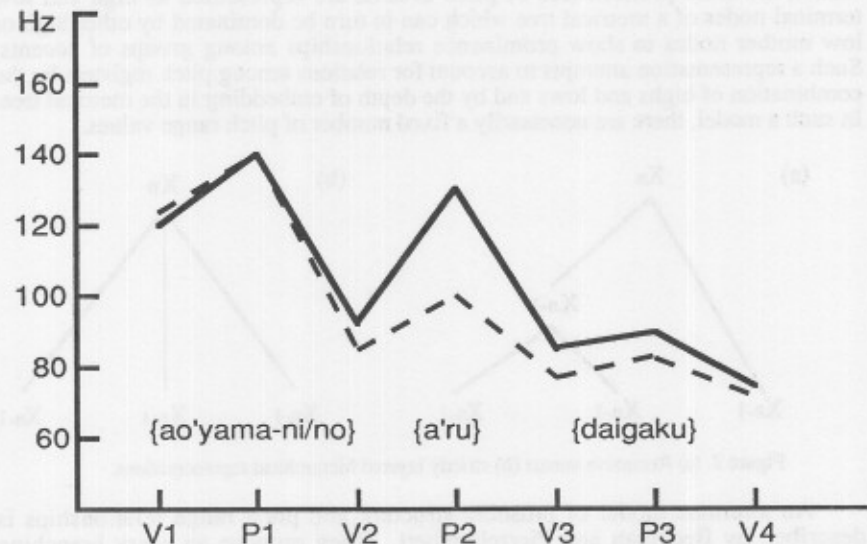


Figure 3. A schematization of the intonation contour of a left-branching (dotted line) versus right-branching (solid line). (Adapted from Figure 15.4 in Kubozono (1992)).

Kubozono's data show that, while in the left-branching structure ((3a) & (4a)) the F0 peak on the first word (Peak 1) is a great deal higher than that on the second

(Peak 2), the right-branching structure ((3b) & (4b)) yields a pattern in which the height of Peak 2 is about the same as that of Peak 1. This pattern of the relative heights between the first two peaks is similar to those observed in the studies by Uyeno et al. and Azuma and Tsukuma described above. Figure 3 is a schematization of Kubozono's mean peak (P) and valley (V) measurements for many tokens of the left-branching (LB) and right-branching (RB) structures in (3). Kubozono not only notes a distinct physical difference in the contours for the two branching structures, but uses these observations to motivate a prosodic representation of utterances such as these. He proposes two contrasting prosodic structures, shown in Figure 4a & b. By using a recursive structure in which minor phrases (mp) are embedded within minor phrases, Kubozono is essentially encoding the difference in syntactic branching directly into the phonological representation of these noun phrases. As I will describe below, he claims that it is necessary for the phonology to be able to access the syntactic structure in order to describe the differing peak height relationships which are observed (cf. Figure 3).

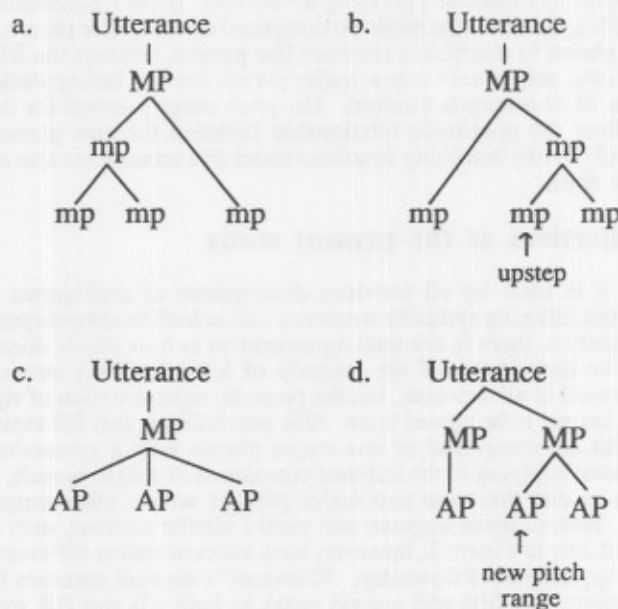



Figure 4. Prosodic structure of (a) left-branching and (b) right-branching noun phrases as proposed by Kubozono (Adapted from example (13) in Kubozono (1992)), and possible (c) left-branching and (d) right-branching structures according to the Beckman and Pierrehumbert model.

By comparing utterances with accented and unaccented initial words, Kubozono (1988, 1989, 1992) argues that downstep, the reduction in pitch range triggered by an accent, is indeed occurring between Peak 1 and Peak 2 (minor phrases 1 and 2) in both of the branching structures shown in Figure 3. The application of this phenomenon will cause the height of Peak 2 to be lower relative to Peak 1 in both cases. While this is an adequate description of the observed downward trend in the LB structure, it is clear that in the RB structure the second peak is higher than it would have been if only downstep had applied. Instead of proposing two types of

downstep, Kubozono introduces the mechanism of 'metrical boost' — a local boost in fundamental frequency which applies to the leftmost constituent of a right-branch. He treats a right-branching structure as 'marked' in a predominantly left-branching language like Japanese. Thus, this phonetic upstep mechanism (see

 arrow in Figure 4b) which he proposes is only indirectly sensitive to syntactic branching via the prosody. To recapitulate then, Kubozono claims that downstep is occurring between Peaks 1 and 2 in both branching structures, but there is an additional upstep mechanism of 'metrical boost' applied to the peak in the RB structure which corresponds to the leftmost minor phrase of a right-branch.

The alternative model of prosodic organization proposed by Beckman and Pierrehumbert might provide a different explanation for the pitch range scaling of Peaks 1 and 2 in the noun phrases discussed above. Though they have not made claims about these particular constructions in Japanese, it is reasonable to presume that the different fundamental frequency patterns of the left-branching and right-branching structures shown in Figure 3 might be accounted for by a difference in the major (or 'intermediate') phrasing for the two. In the LB structure, such as that in (3a) & (4a), the utterance might be composed of one major phrase, within which downstep chains to resemble a staircase like pattern, whereas the RB structure, as in (3b) & (4b), might have a new major phrase starting before Peak 2 so that the application of downstep is blocked. The pitch range selected for the new phrase would reflect the pragmatic relationship between the two phrases. Prosodic structures of the two branching structures under this account are also schematized in Figure 4 (c & d).

## 2.6 Objectives of the present study

Although it is clear by all previous descriptions of ambiguous utterances in Japanese that differing syntactic structures can indeed be disambiguated by means of the intonation, there is not total agreement as to *how* this is done, as was seen above. The description of the prosody of left-branching structures is fairly straightforward in all accounts, but the prosodic representation of right-branching structures has yet to be agreed upon. One possibility is that RB structures such as (3b) & (4b) are comprised of one major phrase with a syntactically sensitive metrical boost applying to the leftmost constituent of a right branch, while another possibility is that there are two major phrases with a pitch range reset at the boundary. Both of these accounts can yield a similar contour, such as that shown by the solid line in Figure 3, however, each account makes different assumptions about the application of downstep. Kubozono's account assumes that downstep occurs between the first and second peaks in both LB and RB structures. The model of Beckman and Pierrehumbert, in contrast, would say that downstep between Peak 1 and Peak 2 will be blocked in the RB structure. Whether there is downstep or not is an empirical issue which deserves to be examined more closely. In addition, since syntactic branching structure is deeply woven into Kubozono's recursive prosodic representation (which in turn determines factors like downstep and metrical boost), a major prediction of his account is that there is only one possible way for speakers to produce RB constructions. Beckman and Pierrehumbert's account does not necessarily predict the lack of consistency, but it is less constrained, in that it can potentially allow for more variation in relative peak heights depending on pragmatic influences. Therefore, it will be of interest to know whether speakers are consistent in their productions of the RB utterances.

Also, Kubozono's discussion is limited to the representation of simple branching structures such as those in the noun phrases. Since similar phenomenon have been noted in ambiguous strings involving more complex constructions such

as relative clauses as well (cf. (1)), we would want a model which would account for these cases in a similar way. Beckman and Pierrehumbert might describe the more complex constructions in a similar way, making 'right-branching' relative clauses like that in (1b) analogous to the right-branching noun phrases, which contain two major phrases. Kubozono, on the other hand, suggests that metrical boost is an n-ary process which will apply multiply according to the depth of the right-branching structure, thus accounting for the larger pitch rise in more complex structures with deeper embedding. In order to give a concrete general account of the prosodic representation of these structures, however, the patterning of downstep must be examined in detail for both noun phrases as well as more complex structures.

This paper discusses the results of two experiments designed to compare these two models of prosodic structure. Both ambiguous noun phrases and relative clause constructions involving an adverb (or locative adjunct) with ambiguous scope of modification were elicited from native speakers of the Tokyo dialect. Fundamental frequency contours were extracted and peak measurements were made in order to examine the behavior of downstep in each construction. Perception experiments for each corpus were also done to confirm the production results. Results show that for the RB noun phrases, in which the syntactic boundary of the right-branching element is not so deep, two of the three speakers exhibited a pattern of downstepping between the first two peaks, as predicted by Kubozono's model. The third speaker showed a downstep reset on the second peak, in accordance with the model proposed by Beckman and Pierrehumbert. This lack of consistency may be due to different speaker strategies for disambiguation. In contrast, for the relative clause constructions, in the right-branching structure (cf. (1b)) whose leftmost edge is deeply embedded, the results for all speakers favored the analysis whereby downstepping is blocked on Peak 2 and the pitch range is reset in the new major phrase. Also, a pause was present in each right-branching structure between Peaks 1 and 2 to aid in the disambiguation. These results suggest that individual speaker strategies or inter-speaker variability as well as depth of embedding are important considerations that must be addressed when proposing a general account of the prosodic structure of Japanese and its interaction with the syntax.

### 3 The experiments

#### 3.1 Experiment 1 — Noun phrases

##### 3.1.1 Production

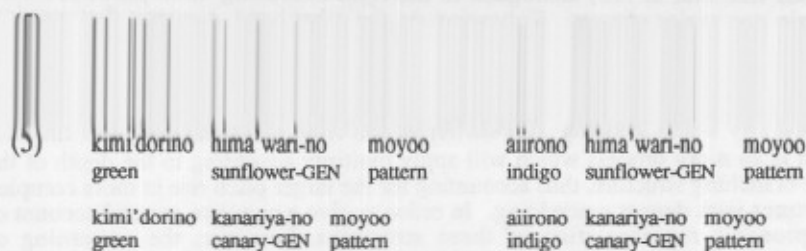
###### Rationale

This experiment was conducted in an attempt to replicate Kubozono's (1988) findings that downstep occurs between Peaks 1 and 2 in both left-branching and right-branching noun phrases. In contrast, Beckman and Pierrehumbert's model suggests that, while downstep will apply between these two peaks in the LB structure, it will be blocked in the RB structure.

###### Methods

Three native speakers of Tokyo Japanese were recorded in a double-walled sound booth at the OSU Linguistics Laboratory. The task in this experiment was to describe various pictures which were mounted on a wall in the booth. The experimenter would point to a picture while saying the prompt *kore-wa nan desu ka?* 'What is this?', and the speaker would then respond using the appropriate noun

phrase in the carrier sentence *sore-wa .... desu* 'That is ....'. The pictures depicted differing interpretations of ambiguous segmental strings. The noun phrases used are given in (5).



Each of the noun phrases was controlled for phonological length (number of morae) of its components as well as vowel height on the relevant morae. This allowed for a direct comparison of peak height. All combinations of accentedness for the first and second words (i.e. +A+A, +A-A, -A+A, -A-A) were included. Each of the four noun phrases is ambiguous in that it has two possible interpretations depending on the branching structure. The left-branching structure would be interpreted as 'the pattern of green / indigo sunflowers', and the right-branching would be 'the green / indigo pattern of sunflowers'. In a preparation session, the speakers were given a concrete context in which these noun phrases might appear, and were informed of the ambiguity involved. They were then asked to describe unambiguously the scene depicted in the picture (there were two pictures per single noun phrase). Ten tokens were elicited for each interpretation for all of the noun phrases for all speakers, resulting in a total of 240 utterances (3 speakers x 2 branching x 2 word1 accentuations x 2 word2 accentuations x 10 tokens). The utterances were elicited in random order.

The utterances were then digitized at 10KHz (12 bit resolution) and the fundamental frequency contour extracted for each token using an autocorrelation-based F0 tracker. The pause duration between offset and onset of voicing of words 1 and 2 was measured, as well as the fundamental frequency value for each peak in the utterance. For utterances in which the first and second words were accented, measurement of the peaks was straightforward. However, in cases where either the first or second word was unaccented, dephrasing often occurred (i.e. the two words were produced in a single accentual phrase), making it impossible to make a 'peak' measurement. Typical contours of the left-branching noun phrases are shown in Figure 5. It is clear from the example contours that there is a tendency for accented words to form their own accentual phrase separate from adjacent words (cf. (a)), while unaccented words tend to be dephrased together with surrounding words (thus not having distinguishable peaks) (cf. (b-d)). Therefore, it becomes impossible to check for downstepping between two peaks, such as Peaks 1 and 2, if they are phrased together. For this reason, the results presented below will only address the *hima wari* type tokens, in which word 2 is accented ((a) & (c)). This assures that there will be a distinguishable Peak 2 whose height can be measured (see discussion of downstep below).

## Results and Discussion

In order to examine the predictions of Kubozono's theory on the one hand, and those of Beckman and Pierrehumbert on the other, it is necessary to look at the downstep patterning in both the left-branching and right-branching interpretations of an ambiguous noun phrase. Example contours of these structures are given in Figure 6.

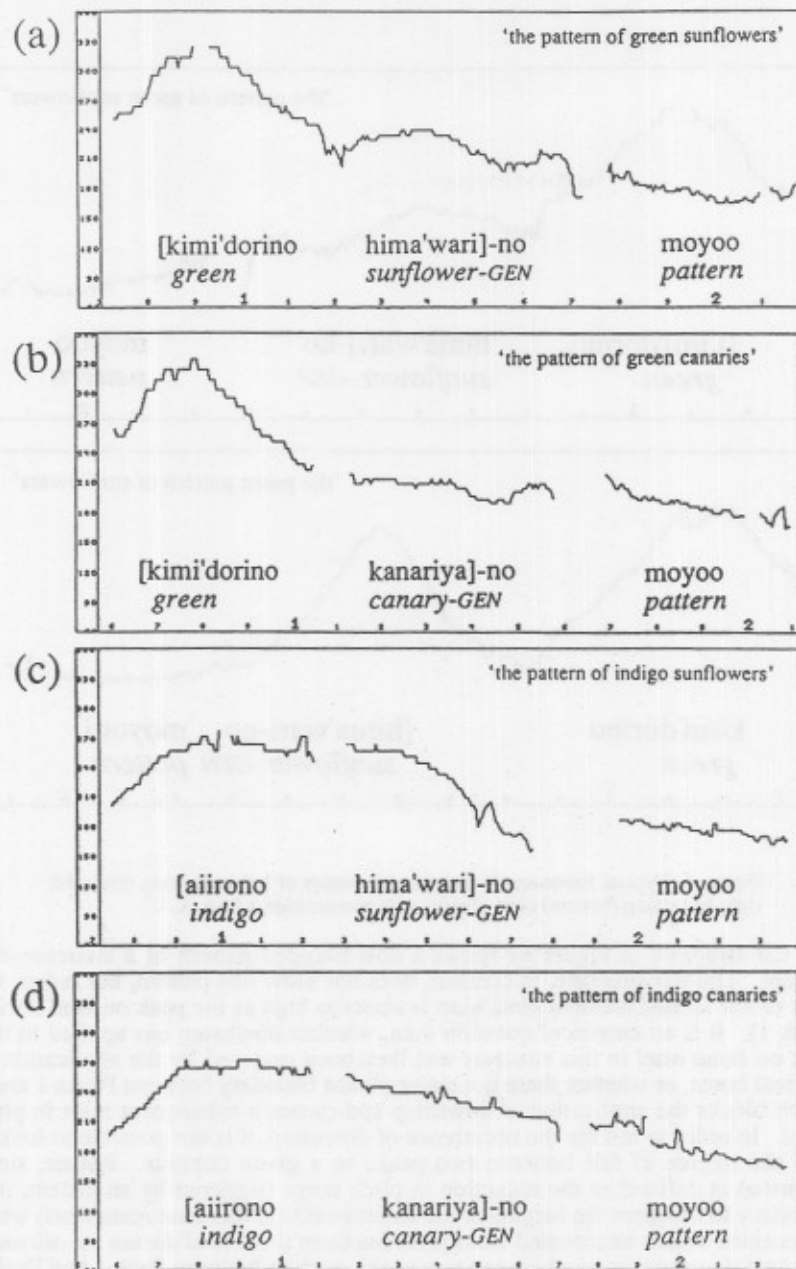


Figure 5. Example contours of left-branching noun phrases. Accentuation of first and second words: (a) +A+A, (b) +A-A, (c) -A+A, (d) -A-A.

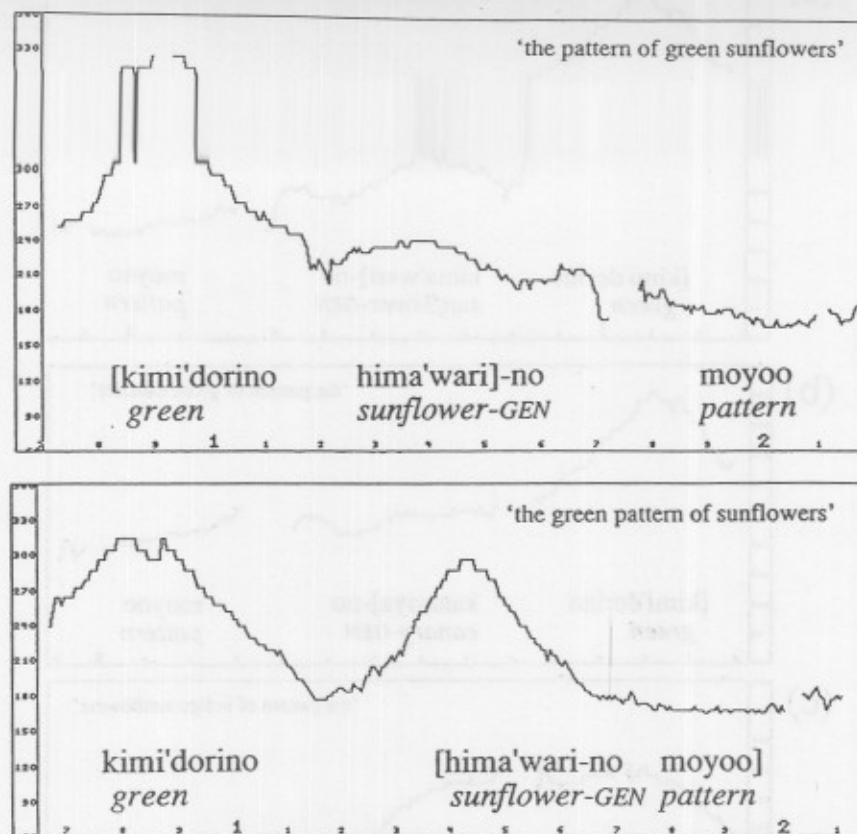


Figure 6. Typical fundamental frequency contours of left-branching (top) and right-branching (bottom) noun phrases with accentuation +A+A-A.

The LB structure in Figure 6a shows a downstepped pattern of a staircase-like descent. The RB structure, in contrast, does not show this pattern, but rather, the peak of the second element *hima'wari* is about as high as the peak on *kimi'dorino* (Peak 1). It is an empirical question then, whether downstep has applied to this peak on *hima'wari* in this structure and then been reversed by the application of metrical boost, or whether there is a major phrase boundary between Peaks 1 and 2 which blocks the application of downstep and causes a subsequent reset in pitch range. In order to test for the occurrence of downstep, it is not possible to look at only the degree of fall between two peaks in a given contour. Rather, since downstep is defined as the reduction in pitch range triggered by an accent, it is necessary to compare the heights of the target peaks (in this case *hima'wari*) when an accented versus unaccented word precedes them (here *kimi'dorino* vs. *aiirono*). If indeed downstep occurs between two peaks, such as between Peak 1 and Peak 2 here, then the prediction is that the height of Peak 2 would be significantly lower when preceded by an accented item than when preceded by an unaccented item. If, on the other hand, there is no downstep occurring between the two peaks, we would expect to see no significant difference in the height of Peak 2 as a function of the accentedness of the preceding word.



Since, as mentioned above, unaccented words tend to phrase together with adjacent words, it becomes impossible to look at the effects of downstep in LB structures (in which dephrasing occurred frequently) — that is, while the accented *kimi'dorino* forms its own accentual phrase whose peak height is readily measurable (cf. Figure 4a), the unaccented *aiirono* phrases together with *hima'wari* (cf. Figure 4c) and thus a 'peak' is not readily measurable. Therefore, I will focus on RB structures only in this examination of downstep between Peaks 1 and 2. Recall that the two contrasting theories of Japanese prosodic organization offer essentially the same account of the LB structures, but differ in their accounts of RB structures.

The graph in Figure 7 shows frequency values of the second peak height (*hima'wari*) when following an accented word *kimi'dorino* versus an unaccented word *aiirono* for one speaker (AM).

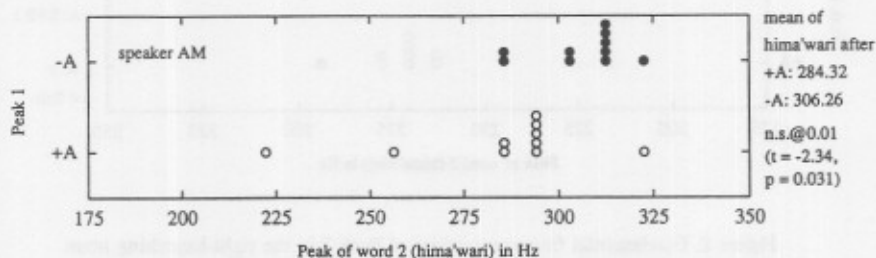


Figure 7. Fundamental frequency values of Peak 2 in the right-branching noun phrase [*kimi'dorino* / *aiirono* [*hima'wari-no moyoo*]] 'the green / indigo pattern of sunflowers'. Full circles indicate word 1 is -A, hollow circles indicate word 1 is +A. Speaker AM.

This graph shows that, with the exception of two outliers (at 225 Hz & 260 Hz) which are clearly downstepped (Peak 2 is quite low), the majority of the tokens following the accented Peak 1 are not substantially lower than those following unaccented Peak 1. Although the average values differ by about 20 Hz, the results of a t-test analysis indicate that the means are not significantly different ( $t = -2.34$ ,  $p > 0.01$ ).<sup>1</sup> The two outliers may be taken to be of a different population from the rest of the tokens, in which case a t-test using all tokens is rendered inappropriate. However, even a comparison of the values excluding the two outliers shows that the samples are not significantly different ( $t = -1.89$ ,  $p > 0.01$ ). These results indicate that, for speaker AM, there is no downstep occurring between Peaks 1 and 2 in the majority of the cases, disputing Kubozono's claim that it actually does apply in such constructions. This absence of downstep would be interpreted within a framework like that proposed by Beckman and Pierrehumbert as an indication of the presence of a major phrase boundary between these two peaks. Therefore, it appears that for the majority of utterances of the RB structure for this speaker, she in fact produced two major phrases, across which downstep is blocked (cf. Figure 4d).

However, for the two other Tokyo speakers used in this study, the patterning of downstep resembles the findings of Kubozono. Figure 8 shows the same comparison of the height of Peak 2 when following an accented versus unaccented word, as shown for speaker AM above.

<sup>1</sup>The 0.01 level of significance was used in all of the t-tests in this study.

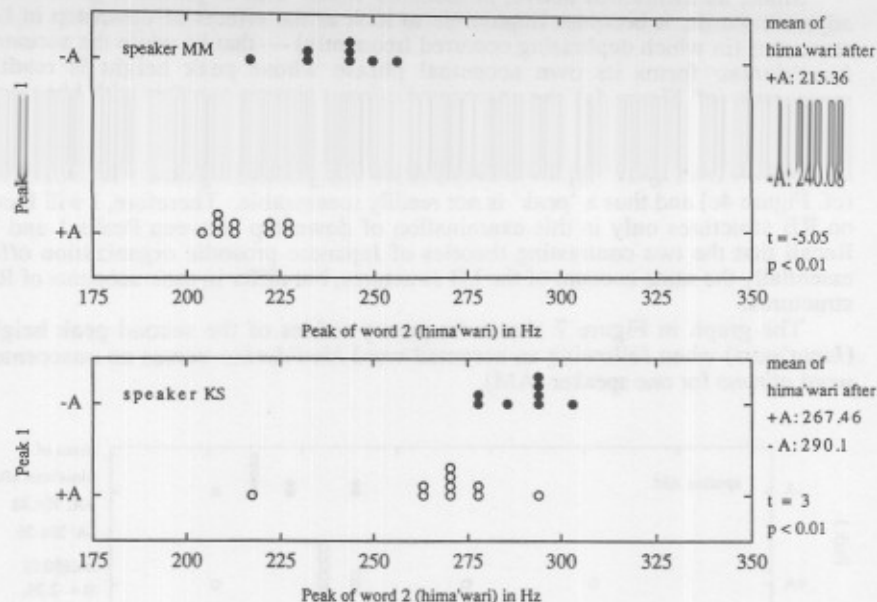


Figure 8. Fundamental frequency values of Peak 2 in the right-branching noun phrase [kimi'dorino / aiirono [hima'wari-no moyoo]] 'the green / indigo pattern of sunflowers'. Full circles indicate word 1 is -A, hollow circles indicate word 1 is +A. Speakers MM and KS.

In contrast to speaker AM, both speakers MM and KS showed a significant downstep relationship between Peaks 1 and 2 even in this RB structure. As can be seen from Figure 8, the height of Peak 2 when following an accented word is significantly lower (MM:  $t = -5.05$ ,  $p < 0.01$ ; KS:  $t = 3.0$ ,  $p < 0.01$ ) than when it is following an unaccented word. This is true for the data as a whole — they include no obvious outliers which we could say are phrased differently. This pattern of downstepping between Peaks 1 and 2 replicates the findings by Kubozono (1988, 1989, 1992), and supports his claim that the only difference between LB and RB structures is the presence or absence of metrical boost on the already downstepped leftmost constituent of a right branch.

Measurement of pause duration shows that, while all tokens of the left-branching structure were uttered with no pause between Peaks 1 and 2, the right-branching structures were equally divided as to whether there was a pause at this location or not. This indicates that the presence or absence of a pause alone cannot disambiguate the structures, but that indeed the contribution of the F0 contour plays a major role. This result supports Azuma and Tsukuma's (1990, 1991) findings that F0 is a more salient factor in disambiguation. Further discussion of the role of pauses in the present data will be addressed in the perception section 3.1.2.

This experiment involving ambiguous noun phrases was conducted in an attempt to replicate Kubozono's (1988) findings that downstep does indeed occur between the first and second peaks in right-branching constructions, a discovery which led him to propose the syntactically induced upstep mechanism of metrical boost. These results were replicated for two of the three Tokyo speakers used in the experiment, while the third speaker showed a behavior closer to that described by the Beckman and Pierrehumbert model. This suggests that, in the disambiguation of noun phrases, individual speakers may make use of different strategies,

involving variant prosodic phrasings or optional application of metrical boost. This variation suggests that it may not be appropriate to explain the 'boost' in peak height on the second peak in RB structures as something that is necessarily tied to the syntactic configuration (as is the case with metrical boost), but rather it can be thought of as an increased local pitch prominence and/or phrase break used by speakers to signal the disjuncture between the two adjacent words.

### 3.1.2 Perception

#### Rationale

In accordance with the proposals set forth by Uyeno et al. (1980) and Azuma and Tsukuma (1990, 1991), in which the height of the second peak compared to the first influenced the listeners' interpretation of the ambiguous sentence, it was predicted that in this perception experiment as well, there would be a similar correlation between the difference in height between Peaks 1 and 2 and listener judgment. Specifically, it was hypothesized that a large positive difference (Peak 1 is substantially higher than Peak 2) would cue the A interpretation (LB), while a negative difference or no difference (Peak 2 higher than or equal to Peak 1) would cue the B interpretation (RB). If this holds true, we should observe a positive linear correlation between the listener choice and the difference in height of the peaks.

#### Methods

Twelve native listeners of Tokyo Japanese participated in this experiment. Five tokens of each of interpretations A (left-branching) and B (right-branching) produced by each speaker were selected randomly from the *kimi'dorino hima'wari* and *kimi'dorino kanariya* types. This gave a total of 60 utterances (3 speakers x 2 branching x 5 tokens x 2 word2 accentuations). Each token was presented to the listener twice (randomly), making a total of 120 stimuli. The stimuli were transferred from digitized form onto an audio tape, and were presented in blocks of 15 utterances each, with the type of the token being constant within a block. The stimuli within each block were randomized, and played at 5 second intervals. Each occurrence of a token was heard only once. The subjects listened to the audio stimuli over headphones in a double-walled sound booth, with the relevant visual cues used in the production experiment (labeled A and B) mounted on a wall directly in front of them. Subjects were asked to listen to each utterance and judge which picture it was intended to describe on a five point scale from 'definitely interpretation A' through 'I don't know' to 'definitely interpretation B'. These choices were explained to the listeners before the start of the experiment by an instruction sheet written in Japanese. The pictures corresponding to A and B were reversed after half of the listeners had taken the test in order to avoid response bias due to order of responses. The participants were informed of the ambiguity of the sentences in a practice session beforehand and were asked to think of how they themselves might describe each picture unambiguously. No verbal prompt was given. There was a sample block prior to the actual test which was repeated as many times as needed in order for the listeners to feel comfortable with the task. At the close of the session, each participant was asked what cues they listened for in attempting to distinguish interpretation A from interpretation B. Listener judgments ranged from 54% (near chance level) to 95% correct.

#### Results

Figure 9 shows the listener choice (sum of choices for all listeners), ranging from 100% A judgments to 100% B judgments, plotted against the difference in ERB

(equivalent rectangular bandwidth) between Peaks 1 and 2 in each token. The ERB psychoacoustic measure was chosen since it has been shown to best reflect the scaling of pitch prominences in the perception of intonation (Hermes & van Gestel, 1991; Moore & Glasberg, 1983).<sup>2</sup> Qualitatively similar results can be found when

substituting the difference in Hz or semitones for this measure.

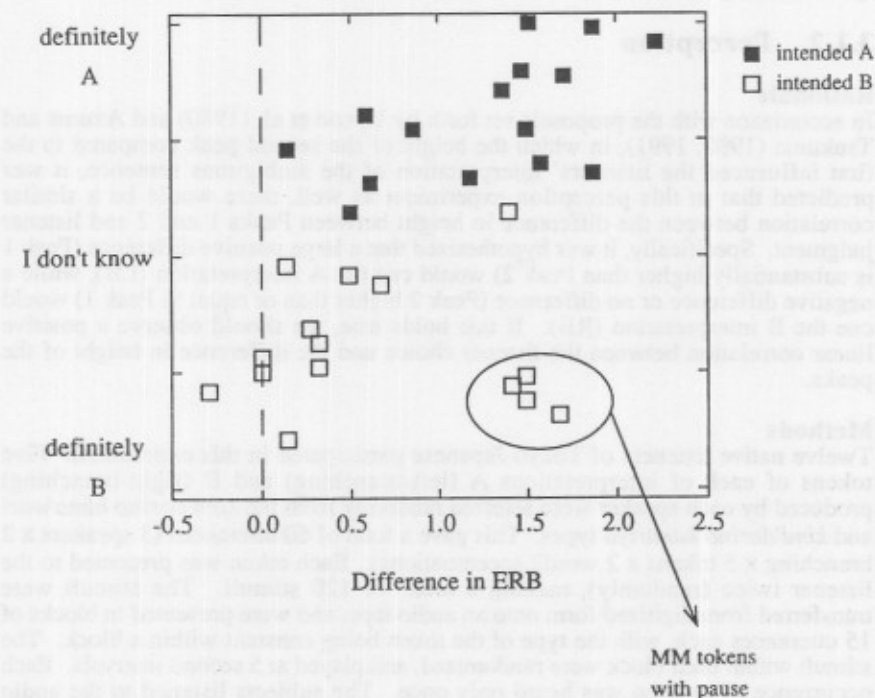


Figure 9. Listener choice from 'definitely A' to 'definitely B' plotted as a function of difference in ERB between Peaks 1 and 2 for [kimi'dorino hima'wari-no moyoo] noun phrases.

The data points in this graph show a gradual transition from 100% 'definitely A' judgments in the upper right-hand corner (greatest difference in ERB) to 100% 'definitely B' judgments in the lower left-hand corner (negative or no difference in ERB). These results support the hypothesis that the difference in the height of the peaks is directly correlated with the listener choice. The only apparent exceptions to this overall trend are the four outliers of the intended B type clustered around 1.5 ERB. Closer examination of these outliers shows an interesting effect concerning the role of the pause in judging these utterances. Each of these outlying tokens was produced by speaker MM, who made Peak 2 subordinate to Peak 1 (large positive difference in ERB) even in the right-branching constructions. The other speakers did not employ this strategy, though speaker MM seemed to use it often. The graph in Figure 9 shows that these outliers were judged by the native listeners as close to

<sup>2</sup>The equation used to calculate ERB-rate (equivalent rectangular bandwidth) is:  $E = 11.17 \ln \left( \frac{f+0.312}{f+14.675} \right) + 43.0$  where  $f$  is in kHz (Moore and Glasberg, 1983).



A-like by the majority of listeners. This indicates that there are other factors beside the pause which are influencing judgments; namely, the difference in peak heights shown in Figure 9. In the short interview after the experiment, listeners were asked which cues they listened for in differentiating the two interpretations. The



(6)	attended to pause (for B) only:	7 listeners
	attended to peak height difference only:	1 listener
	attended to both:	3 listeners
	neither:	1 listener

This suggests that the pause was the most salient cue to disambiguation. However, these impressions may not be too reliable, since there were many tokens which did not have a pause but still were correctly judged as type B tokens. It may be that the listeners were not conscious of all the cues involved, and just named the first which came to mind or the ones that they could most easily describe in words.<sup>3</sup>

In conclusion, results of this perception study show that while the difference in height of Peaks 1 and 2 is strongly correlated to the listener judgment, the presence or absence of a pause also plays a non-trivial part in disambiguation. This supports previous claims (e.g. Lehiste et al, 1976) that native listeners can take into consideration multiple prosodic factors in the perception of differing syntactic structures.

## 3.2 Experiment 2 — Relative clause constructions

### 3.2.1 Production

#### Rationale

In order to have a general account of the syntax-prosody relation for differing branching structures, it is necessary to examine structures which involve more complex branching than just LB and RB noun phrases. This experiment was designed to compare the accounts of the two alternate theories of Japanese prosodic organization for complex structures such as relative clause constructions with ambiguous scope of modification of temporal or locative adjuncts. Again, examination of downstepping patterns will be relevant for the assessment of the two accounts.

#### Methods

The three native speakers of Tokyo Japanese who participated in experiment 1 participated in this experiment as well. The task in this case was to read aloud a randomized list of sentences. The corpus is given below in (7).

Each sentence here is ambiguous in the sense that the initial adverb (or locative adjunct) can be taken to modify the verb of the relative clause directly following it (interpretation A: 'The scarf that I knitted last year was stolen. '), or it can modify the verb of the matrix clause (interpretation B: 'The scarf that I knitted was stolen last year. ').

<sup>3</sup> Mineharu Nakayama has also suggested that the relatively higher peak height for Peak 2 in the RB structure may give listeners the impression of a pause. This is certainly a possibility, since it has been documented in English at least that adjacent peaks with about the same prominence can give the impression of a salient juncture or break.

(7) a. Temporal-*eri'maki* set:

kyo'nen a'nda eri'maki-ga nusuma'reta  
last year knitted scarf-NOM was stolen

yuube a'nda eri'maki-ga nusuma'reta  
last night knitted scarf-NOM was stolen

b. Locative-*eri'maki* set:

Me'jiro-de a'nda eri'maki-ga nusuma'reta  
Mejiro-LOC knitted scarf-NOM was stolen

Ueno-de a'nda eri'maki-ga nusuma'reta  
Ueno-LOC knitted scarf-NOM was stolen

The list of randomized sentences was presented to the speakers in normal Japanese orthography (kanji and kana), with a cue beneath indicating which meaning they should utter, as exemplified in (8). The speakers were asked to produce each sentence with the meaning indicated in the cue.

- (8) yuube a'nda eri'maki-ga nusuma'reta.  
(yuube a'nda) 'You knitted it last night.'

[actual cues were of course written in Japanese orthography without accompanying English translation.]

The speakers were given instructions for the task beforehand, presented in Japanese orthography, which outlined the ambiguity of the sentences and the contexts in which each interpretation might be uttered. Each speaker was allowed to practice before being recorded, but no verbal feedback or prompts were given. Ten tokens of each interpretation of all the utterances were elicited from each speaker, totaling 240 utterances (3 speakers x 2 branching x 2 type adjuncts x 2 word1 accentuations x 10 tokens). The tokens were analyzed as outlined in §3.1.1 above.

### Results and discussion

The results from the two sets are essentially the same, as predicted from the fact that the accentuation of the components are identical, the only difference between the two is the type (temporal or locative) of adjunct being adjoined to S.<sup>4</sup> In light of this similarity, the discussion below will focus on the temporal-*eri'maki* set.

Figure 11 shows sample contours for both interpretations. The results for this experiment showed more consistency across the three speakers than did the results for the noun phrase case. As with the previous experiment, the behavior of the peak following accented versus unaccented first phrases was examined in order to determine whether or not there exists a downstep-blocking major phrase break between Peaks 1 and 2 in the right-branching structure. In contrast to the noun phrases, it may not be appropriate to characterize these relative clause constructions as 'left-branching' or 'right-branching', since relative clauses in Japanese are all left-branching. The distinction to be made between the two structures here is characterized by Uyeno et al. (1980) as 'left-branching', in which the initial adverb (or locative adjunct) modifies the verb of the relative clause, versus 'center-embedding', in which the adverb modifies the verb of the matrix clause. As far as

<sup>4</sup>The syntactic structures of these sentences are given in Figure 18 below.

pitch range relationships are concerned, these two structures behave very similarly to LB and RB noun phrases, respectively. In the following description, I will refer to the left-branching relative clause as interpretation A, and the center-embedding relative clause as interpretation B.

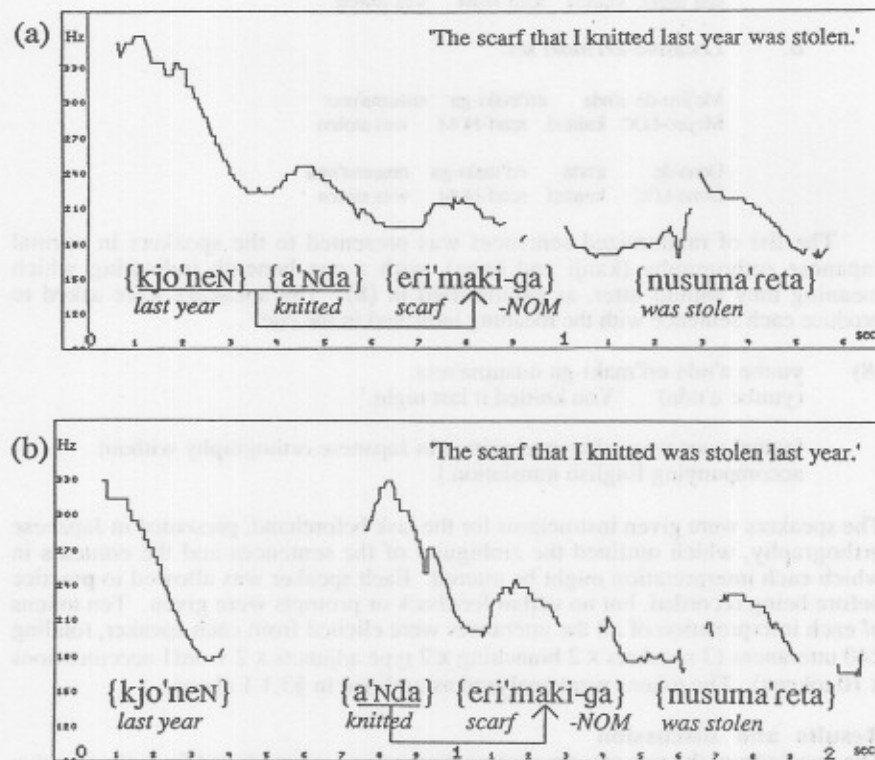


Figure 11. Typical fundamental frequency contours of interpretations A (top) and B (bottom) of the ambiguous string [kjo'nen a'nda eri'maki-ga nusuma'reta].

In order to examine the downstep relationship between Peak 1 (initial adverb or locative adjunct) and Peak 2 (verb of the relative clause), plots of the frequency of Peak 2 when following an accented and unaccented initial word are shown in Figure 12, for one speaker (KS). Again, as with the noun phrases, because of the dephrasing of the initial unaccented words in the left-branching (A) interpretations, only the relationships of the peaks in the B type (analogous to the RB structure) could be examined.

In both sets it is clear that the height of Peak 2 (*a'nda*) is not significantly lower when following an accented word as opposed to an unaccented word (adverb:  $t = -2.28$ ,  $p > 0.01$ ; locative:  $t = -1.51$ ,  $p > 0.01$ ). This result was found for the other two speakers as well, for both sets of utterances. Thus, according to the criteria for the detection of downstep set by Poser (1984) and confirmed by Pierrehumbert and Beckman (1988), Kubozono (1988), and others, there does not appear to be any



downstep occurring between these two peaks. Rather, its application is blocked, and the pitch range is reset on Peak 2 at the start of a new major phrase. These results support the account by Beckman and Pierrehumbert as well as the claims of Selkirk and Tateishi (1991) who propose a major phrase break for these types of structures. They form an apparent contradiction to Kubozono's claim that metrical boost is applied to a downstepped peak to create the pitch range expansion in complex structures such as these.

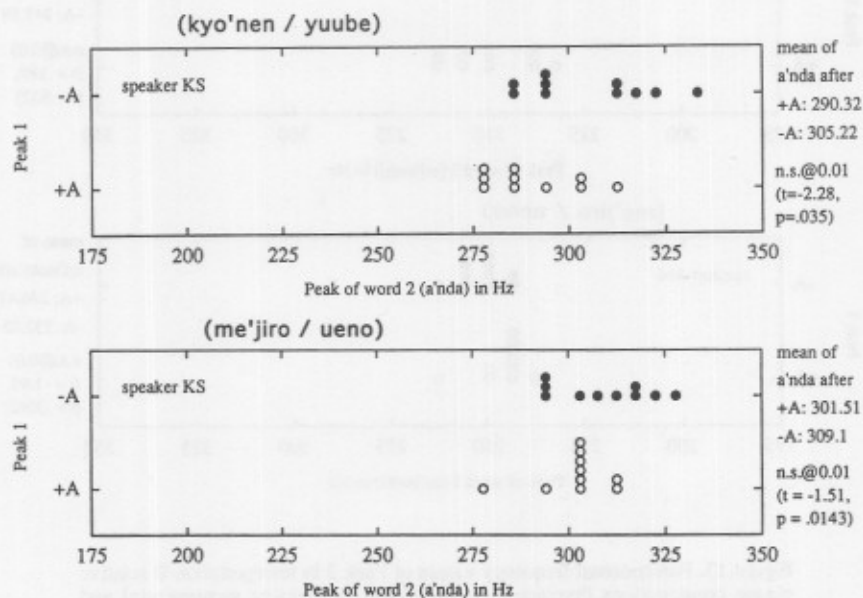


Figure 12. Fundamental frequency values of Peak 2 in interpretation B relative clause constructions [kyo'nen / yuube [a'nda] eri'maki-ga nusuma'reta] and [Me'jiro-de / Ueno-de [a'nda] eri'maki-ga nusuma'reta]. Filled circles indicate word 1 is -A, hollow circles indicate word 1 is +A. Speaker KS.

Since downstep is said to apply iteratively within the major phrase, it is useful to look at the relationship between the height of Peak 1 and Peak 3. If downstep is indeed blocked between Peaks 1 and 2, as the data seem to indicate, then one would expect that the accentedness of Peak 1 would also not have an effect on the height of Peak 3 (*eri'maki*). If, on the other hand, downstep does apply between Peaks 1 and 2, then it is expected that the difference in the height of Peak 3, even in the type B relative clauses, would show some signs of downstep chaining onto it. The result of a comparison of Peaks 1 and 3 is shown in Figure 13 for speaker AM, a representative case.

It is clear from Figure 13 that there is no effect of downstep between Peaks 1 and 2 which may be chaining onto Peak 3 in the B type structure (adverb:  $t = .189$ ,  $p > 0.01$ ; locative:  $t = -1.93$ ,  $p > 0.01$ ). This result holds true for the two other speakers as well: for all speakers, the height of Peak 3 was not lower when following an initial accented word as opposed to an initial unaccented word. This can be described in Beckman and Pierrehumbert's framework by saying that there

is an major phrase boundary which blocks downstep between Peaks 1 and 2 in these type B constructions.

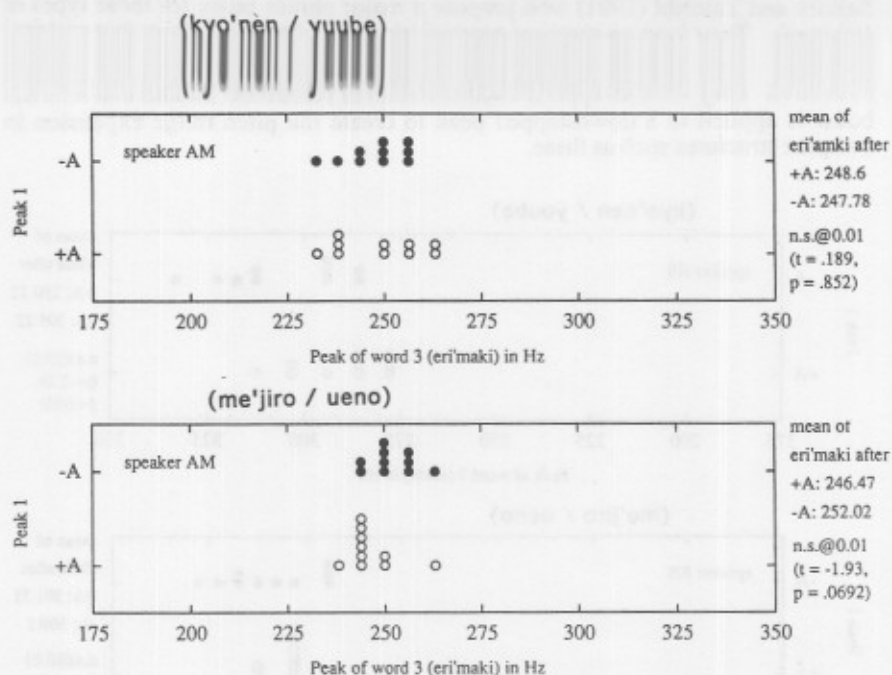


Figure 13. Fundamental frequency values of Peak 3 in interpretation B relative clause constructions [kyo'nen / yuube [a'nda] eri'maki-ga nusuma'reta] and [Me'jiro-de / Ueno-de [a'nda] eri'maki-ga nusuma'reta]. Filled circles indicate word 1 is -A, hollow circles indicate word 1 is +A. Speaker AM.

Another interesting result of this study concerns the scaling of initial peak height. Figure 14 shows frequency distributions of initial accented and unaccented peak heights for both branching structures. (Here, both the Temporal-*eri'maki* and Locative-*eri'maki* sets have been combined.) These distributions show that, for both accented and unaccented initial words, the height of the initial peak is significantly higher in type A constructions than it is in the type B constructions (-A:  $t = 8.33$ ,  $p < 0.01$ ; +A:  $t = 8.12$ ,  $p < 0.01$ ). These results hold true for the two other speakers as well, with the exception of the accented initial words for speaker KS, where the difference only approaches significance at the 0.01 level (MM-A:  $t = 10.8$ ,  $p < 0.01$ ; MM+A:  $t = 6.63$ ,  $p < 0.01$ ; KS-A:  $t = 3.4$ ,  $p < 0.01$ ; KS+A:  $t = 2.62$ ,  $p < 0.05$ ).

Such findings are interesting since they suggest that the speaker needs to look ahead at the syntactic configuration of an utterance even before s/he actually utters the first word. The overall 'mental plan' of the sentence will thus effect the pitch range scaling of even the leftmost component. If the initial adjunct modifies the immediately following verb, the speaker will utter it with a higher overall pitch than if it modifies the matrix verb three words later. This variability in initial peak scaling according to the syntactic structure of the rest of the utterance is similar to the findings of Ladd for English (Ladd, 1988; Ladd & Johnson, 1987). Ladd

states that "it is both necessary and appropriate to enrich the phonological representation of intonation in order to express the fact that syntactic organisation may be signalled intonationally in fine differences of peak height." (Ladd, 1986: 329) It was this notion which led him to propose his metrical representation of pitch range, in which the syntactic structure is encoded into an overall phonological 'plan' of the utterance, as was briefly outlined in §2.4 above. Ladd proposes that downstep can be modeled by a high-low branching node as in (9).

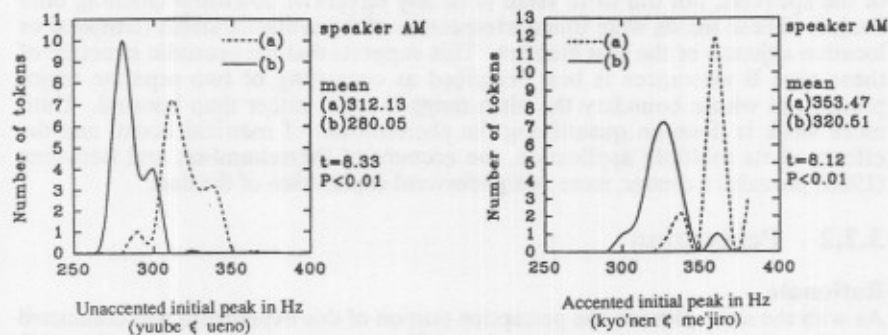


Figure 14. Frequency distributions of initial peak height in type A (dotted line) and type B (solid line) relative clause constructions. Temporal-*eri'maki* and Locative-*eri'maki* sets combined. Speaker AM.

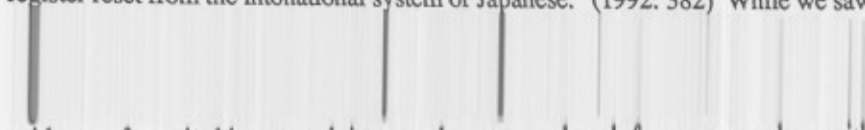
(9)



The metrical tree of an utterance will mimic the syntactic structure by the direction of branching or the depth of embedding. The pitch register relationships which are in effect a consequence of this structure are represented by different configurations of the downstepping h-l sequence, or the non-downstepping l-h. Such a model is attractive since it by definition encodes pitch register relationships into its representation. In a model such as the one proposed by Beckman and Pierrehumbert, there is no way of predicting the relative heights of the first and second words in the two structures directly from the prosodic representation itself. One would have to rely on a discourse structure or some other syntactically sensitive structure to provide the information of relative pitch heights. Thus, Ladd's model does seem attractive, if one wants to encode the notions of syntactic branching into the phonological representation. However, his model is only one of relative pitch relationships, and until enough quantitative data are presented and modeled to determine exactly how to translate these 'h's and 'l's into actual pitch values, the proposal still remains speculative.

This experiment involving ambiguous scope of adverbial (or locative adjunct) modification in relative clause constructions was carried out in an attempt to examine Kubozono's conjectures about the multiple application of metrical boost in deeply embedded right-branching structures. Kubozono (1989, 1992) suggests

that the same notion of metrical boost seen applying to already downstepped peaks in right-branching noun phrases can be expanded to explain any boost in pitch range at the left edge of a right syntactic branch. He proposes that, with metrical boost as an n-ary process, it is possible "to eliminate the conventional rule of pitch register reset from the intonational system of Japanese." (1992: 382) While we saw



evidence of metrical boost applying to a downstepped peak for some speakers with noun phrases, it is not clear that this is an appropriate characterization of the more complex utterances. Results from the second experiment showed that there was in fact no downstep occurring between Peaks 1 and 2 in the type B utterances for any of the speakers, nor did there seem to be any effects of downstep chaining onto Peak 3. These trends were found irrespective of the syntactic status (temporal or locative adjunct) of the first element. This suggests that the prosodic structure of these type B utterances is best described as consisting of two separate major phrases, at whose boundary the pitch range is reset, rather than boosted. Until more work is done on quantifying the phenomenon of metrical boost, and the effects of its multiple application, the account of Pierrehumbert and Beckman (1988) provides a clearer, more straightforward explanation of the data.

### 3.2.2 Perception

#### Rationale

As with the noun phrases, the perception portion of this experiment was conducted in order to confirm that the production differences between type A and type B structures are perceptually salient. Again, in accordance with the proposals of Uyeno et al. (1980) and Azuma & Tsukuma (1990), we would predict that a large positive difference between the height of the first two peaks is more likely to cue the A interpretation, while a negative, very small positive difference, or no difference will cue the B interpretation.

#### Methods

Eleven native listeners of Tokyo Japanese (or near Standard) participated in this experiment. One speaker was excluded since her results (4% correct) indicated that she could not attend to the task. Essentially the same method as the noun phrase perception experiment was used with these relative clause constructions. In this case, only the utterances in the Temporal-*eri'maki* set were used as stimuli, and every token was presented only once. This gave a total of 120 stimuli in all (3 speakers x 2 branchings x 2 word1 accentuations x 10 tokens). Each of the ten participants had judgments of 86% to 98% correct.

#### Results and discussion

As with the data of the relative clause construction production test, the results of this perception experiment were more consistent than the noun phrase perception. There was better agreement among listeners as to whether the token was the type A or type B interpretation, and there were no outlying tokens. Figure 15 shows listener judgment plotted as a function of the difference in the height of Peaks 1 and 2. Again, due to dephrasing, only the all accented type *kyo'nen a'nda eri'maki-ga nusuma'reta* could be examined.

It is clear from this graph that listeners were able to identify the structure of the utterance (interpretation A or B) with much greater success than they had with the noun phrases. Therefore, since most of the points are clustered around the two extremes, it is difficult to see if there is a transition from 'definitely A' (upper right) to 'definitely B' (lower left) as was clear with the noun phrases. There does seem to be a slight tendency for the A tokens with a smaller ERB value to extend toward

the middle of the graph, showing the traces of a correlation between the difference in height and the listener judgment. Figure 16 shows a similar clear-cut pattern in the relation between pause and the listener choice.

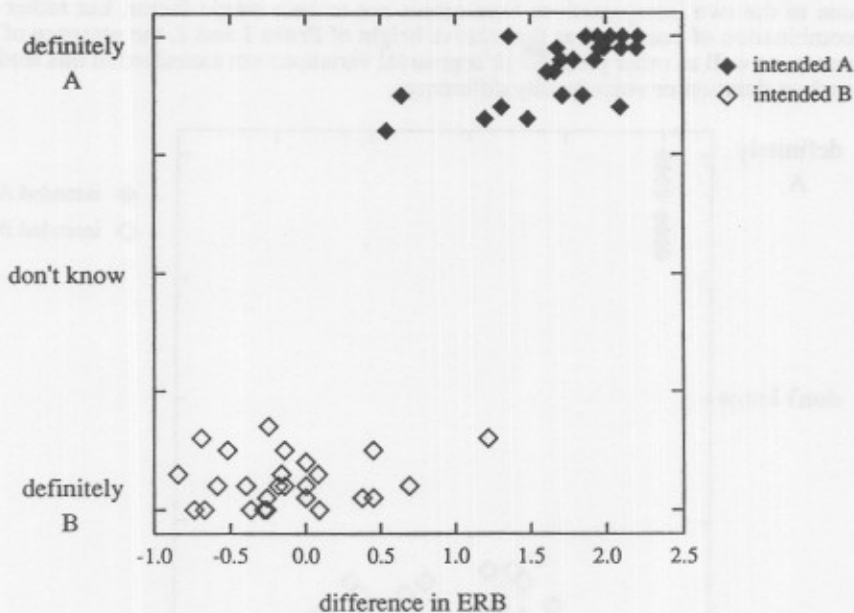


Figure 15. Listener choice from 'definitely A' to 'definitely B' plotted as a function of the difference in ERB between Peaks 1 and 2 for [kyo'nen a'nda eri'maki-ga nusuma'reta] relative clause constructions.

This graph clearly shows that none of the intended A tokens had a pause occurring between Peaks 1 and 2, while all of the intended B tokens contained a pause.<sup>5</sup> It might therefore seem that it is the pause only which distinguishes the two interpretations. However, the fact that the judgments for tokens in both A and B categories vary suggests that there is something in addition to this which influences native listeners' perception of these structures. A table outlining the cues which listeners after the experiment named as being relevant to disambiguation is shown in (10).

- |      |  |             |
|------|--|-------------|
| (10) | attended to pause (for B) only:          | 1 listener  |
|      | attended to peak height difference only: | 1 listener  |
|      | attended to both:                        | 6 listeners |
|      | neither:                                 | 2 listeners |

<sup>5</sup>None of the speakers in this study produced type B utterances without a pause, however, it is possible to do so. Speaker YO in a follow up experiment uttered the same constructions with little or no pause, due to the rapid rate of speech. Also see Uyeno et al. (1980) and Azuma & Tsukuma (1990, 1991) for B type utterances without pauses.

In this perception test, listeners tended to be more aware of the difference in peak heights, and felt that, together with a pause, this could cue the appropriate interpretation. However, as noted before, these impressions may not be a reliable indicator of the actual cues which native speakers listen for.

It is therefore possible to conclude that, as with the noun phrases, the perceptual

cue to the two interpretations here seems not to be a single factor, but rather a combination of cues such as the relative height of Peaks 1 and 2, the presence of a pause, as well as other prosodic or segmental variations not examined in this study such as duration or voice quality differences.

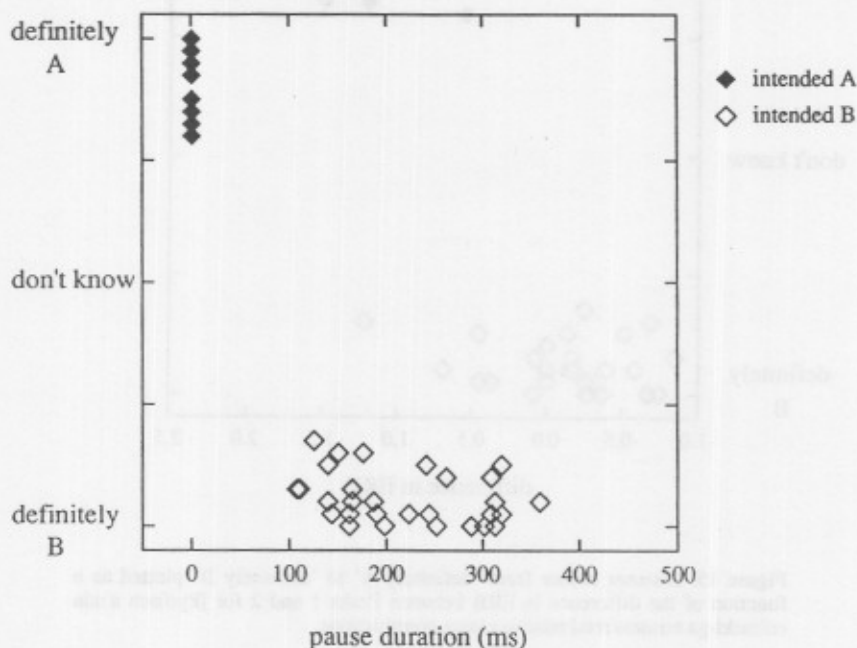


Figure 16. Listener choice from 'definitely A' to 'definitely B' plotted as a function of length of pause between Peaks 1 and 2.

#### 4 Discussion

The present study examined downstep relationships between the first two peaks of ambiguous noun phrases and relative clause constructions. It was carried out in an attempt to replicate Kubozono's (1988) findings for noun phrases that downstep can apply across a syntactic boundary. This study was also designed to test his hypothesis for more complex constructions like relative clauses. The results show that, for noun phrases, two of the three Tokyo speakers behaved as did Kubozono's speaker, downstepping Peak 2 in relation to Peak 1 even in the right-branching structure. However, the third speaker did not show any evidence of downstep between the two peaks in a majority of her utterances, but rather had a major phrase break there which blocked downstep and induced a reset in pitch range. This suggests that the prosodic manifestations cannot be directly attributed

to a certain syntactic branching configuration, and that speakers may use different strategies involving variant prosodic phrasing or optional application of metrical boost. For relative clauses, on the other hand, all speakers used in this study behaved according to the Beckman and Pierrehumbert model. In these structures with a deep syntactic boundary, speakers chose to disambiguate the structures by means of the major phrasing. An examination of the initial peak scaling in these constructions suggests that the choice of the fundamental frequency value on this peak depends on the syntactic branching of the utterance. This resembles other findings by Ladd which originally motivated his metrical representation of pitch register. Perception tests on each of the corpora suggest that listeners use both the difference in the height of Peaks 1 and 2 as well as the presence of a pause as cues to disambiguate the structure of the continuous stream of segmental information.

## 5 Implications for syntax-prosody mapping

It is clear both from previous descriptions of structural ambiguity and from the results of the present study that the prosodic manifestation of an utterance is indeed influenced by its syntactic structure. While this has been taken as a matter of fact by most researchers, less attention has been paid to how exactly this mapping between the syntax and prosody is achieved. The following discussion will outline briefly some previous accounts and algorithms of this mapping, and examine whether they can account for the results of the present study.

One attempt at characterizing the relationship between the syntactic and prosodic structures has already been outlined above in some detail. In this account, proposed by Kubozono (1988, 1989, 1992), the prosody directly accesses certain syntactic configurations — here, the marked left-branching structure. He claims, as was noted above, that an upstep mechanism called 'metrical boost' applies to a minor phrase which finds itself as the leftmost constituent of a right-branch. Consider in Figure 17 the structures of noun phrases such as those discussed in this study (assuming an X' type of syntactic representation).

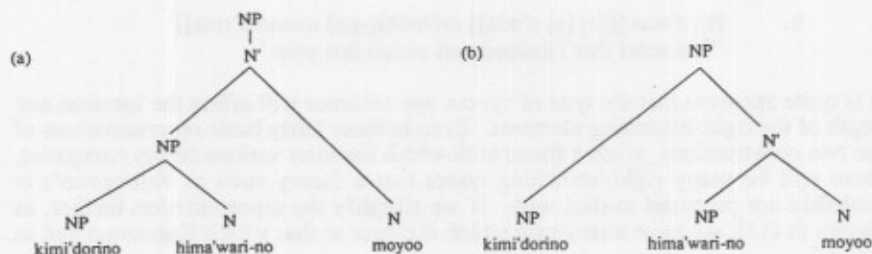


Figure 17. Syntactic structures of left-branching (a) and right-branching (b) noun phrases.

Assuming these structures, the account proposed by Kubozono is straightforward — the minor phrase which falls on the right-branch in (b) will receive a metrical boost to raise its already downstepped peak (cf. Figure 4b). Kubozono chooses to encode the syntactic structure into a recursive prosodic representation which then triggers metrical boost. However, it is also possible to access the syntax directly without encoding it in the phonology first. The end result will be the same.

The question then becomes whether such an account will work for constructions of greater complexity than noun phrases and with deeper levels of embedding. Figure 18 gives the structures of the relative clause constructions used in this study.

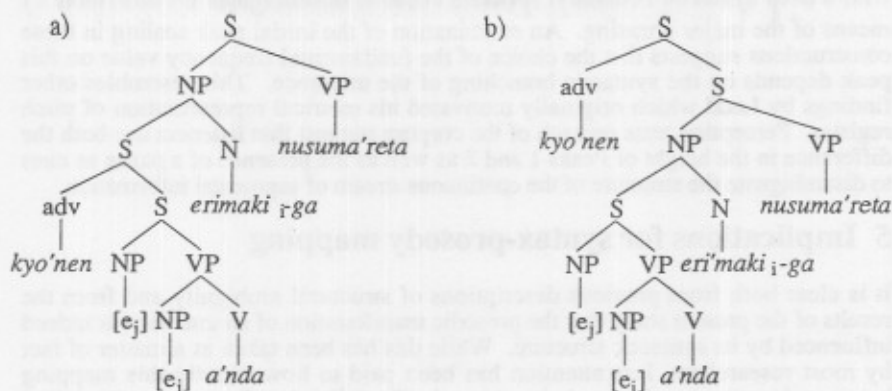


Figure 18. Syntactic structures of type A (a) and type B (b) relative clause constructions.

Kubozono (1992) offers the suggestion that the mechanism of metrical boost is an n-ary process which may apply multiply to right-branching configurations. The structures in Figure 18 are restated in (11) as bracketed strings for easier recognition of the right-branches.

- (11) a. [[[kyo'nen [e<sub>j</sub> [e<sub>i</sub> a'nda]]] eri'maki-ga] nusuma'reta]  
 'The scarf that I knitted last year was stolen.'  
 b. [kyo'nen [[[e<sub>j</sub> [e<sub>i</sub> a'nda]] eri'maki-ga] nusuma'reta]]  
 'The scarf that I knitted was stolen last year.'

It is quite apparent that the type of syntax one assumes will effect the location and depth of the right-branching elements. Even in these fairly basic representations of the two constructions, given a framework which assumes various empty categories, there will be many right-branching nodes that a theory such as Kubozono's is probably not prepared to deal with. If we simplify the representation further, as shown in (12), we have a structure which is closer to that which Kubozono had in mind.<sup>6</sup>

- (12) a. [[[kyo'nen a'nda] eri'maki-ga] nusuma'reta]  
 b. [kyo'nen [[[a'nda] eri'maki-ga] nusuma'reta]]

Applied to such a structure, metrical boost would apply three times on the downstepped verb *a'nda* in (b), causing the peak to be boosted quite a bit. The results discussed above indicate that there is no sign of downstep occurring between Peaks 1 and 2, but rather support an interpretation that the pitch range is being reset completely. While the proposal of a multiple application of metrical

<sup>6</sup>The reader is referred to a similar example in (8b) Kubozono (1992).



boost which in effect 'undoes' downstep totally if applied enough times is still plausible, without more details of the exact quantitative nature of this mechanism and its cumulative effects, it is difficult to support or reject this account.

Other authors have proposed alternative methods of mapping the syntax onto the prosody. Most noteworthy is the edge-based theory proposed by Selkirk (1986) and argued for Japanese by Selkirk and Tateishi (1988, 1991). According to their proposal, the boundary of a major phrase — at which pitch range is reset — corresponds to the left-edge of a maximal projection in the syntactic hierarchy.<sup>7</sup> In such a model, the prosodic structure is influenced by the syntax, but not directly. Selkirk and Tateishi (1988, 1991) have argued that this account can describe the syntactic structures which they examined in their study. It is an empirical question then whether their account will correctly predict the phrasings found in the present study. Given a representation such as that in Figure 17, we would predict that the string would consist of a single major phrase in the LB structure, but two separate major phrases in the RB structure: (kimi'dorino)(hima'wari-no moyoo). This prediction is confirmed by the results of only one speaker in the present study. It fails to explain why we observe an increased prominence on the already downstepped Peak 2 in right-branching structures for the two speakers, who produced the RB strings with just one major phrase. Therefore, in the face of such inter- and intra-speaker variability, such an edge-based mapping theory which relies solely on the syntactic structure will run into trouble.

Still more difficulty for their account arises when we examine the structures of the relative clause constructions (cf. Figure 18). Without going into great detail about how their mapping algorithm might apply to such structures, the crucial thing to note is that, at the level of the most subordinate S, the structures of the two interpretations are identical. Thus, it follows that whatever might apply or not apply to one structure must hold for the other, rendering them virtually identical for such mapping algorithms. Considering only this S, if we assume that phonologically empty nodes are place holders and can serve as the left-edge of a  $X^{\max}$ , then the verb *a'nda* would start a new major phrase in both cases. Likewise, if we choose to ignore empty nodes, then there is nothing on the left-edge of a  $X^{\max}$  in either structure which would predict a pitch range reset.<sup>8</sup> Therefore, the problem that such relative clause constructions with ambiguous scope of adjunct modification hold for such edge-based mapping theories, in virtually any syntactic framework, is that what the algorithm predicts for one structure will be identical to that which it predicts for the other structure.

The last approach to syntax-prosody mapping which I will address here is that of Nespor and Vogel (1986). Theirs is a relation-based theory which holds the notions of head and complement crucial to the relation between syntactic and prosodic structures. They define the phonological phrase and the intonational phrase as the following:

(13) Phonological phrase domain:

Consists of a clitic group which contains a lexical head (X)  
and all clitic groups on its nonrecursive side up to the clitic  
group which contains another head outside of  $X^{\max}$ .  
(paraphrased, Nespor & Vogel, 1986:168)

<sup>7</sup>The edge parameter is set for 'left' in Japanese.

<sup>8</sup>Selkirk and Shen (1990) chose to ignore empty categories, claiming that a phonologically null trace has no effect on the syntax-prosody mapping.

#### Intonational phrase domain:

An intonational phrase may consist of all the phonological phrases in a string that is not structurally attached to the sentence tree at the level of s-structure, or any remaining sequence of adjacent phonological phrases in a root



sentence. (Nespor & Vogel, 1986:189)

First let us consider the relative clause constructions shown in Figure 18. If we take the phonological phrase to be equivalent to the accentual phrase (Vogel, personal communication), and the intonational phrase to be the major phrase, the correct phrasing is predicted. Each of the words in both structures (a) and (b) is a lexical head, and it happens that they are separated from one another by a maximal projection. Therefore, the prediction that each word forms a single accentual (phonological) phrase is correct. However, note that each of these words are accented and thus tend to form their own phrase (cf. §2.2). If one of the words had been unaccented, as in *yuube a'nda eri'maki-ga nusuma'reta*, the prediction would be incorrect since *yuube* is dephrased together with the following verb. Therefore, while the algorithm for determining accentual phrases holds in this structure for accented words, substitution of unaccented words will complicate matters.

Regarding the intonational (major) phrase, the predictions of Nespor and Vogel's mapping algorithm do predict the correct distinction between the two structures in Figure 18. This theory predicts that structure (a) will form one major phrase since all of the accentual (phonological) phrases are part of the root sentence at s-structure, whereas structure (b) would form two major phrases since the adverb is not attached to the root sentence. It is not clear to me if the fact that both the adverb and its sister S are attached to a higher S has any bearing on the validity of their prediction.

However, if we attempt to apply Nespor and Vogel's mapping algorithm to the noun phrase constructions shown in Figure 17, we immediately run into trouble. It cannot account for the speaker variability discussed above, and also has problems with describing dephrasing of unaccented words in the string.

It is obvious from the discussion above that the results of this study present problems for all of the major theories of syntax-prosody mapping: evidence of the fact that the influence syntax has on prosodic structure is anything but straightforward. While Selkirk and Tateishi and Nespor and Vogel's mapping algorithms need to undergo major revision to account for the present data, Kubozono's proposal may be the direction in which we should look when the phenomenon of metrical boost has been more carefully examined and documented.

## 6 Conclusion

Syntactic structure can indeed influence the prosodic realization of an utterance. This has been shown not only in the present experiments but in numerous previous studies of several languages. The more relevant issues are exactly *how* it does this and to what extent we should encode this influence into our phonological representation. The present study looked at structurally ambiguous utterances in Japanese and how disambiguation of these is achieved via the fundamental frequency contour. Structures involving left-branching nodes are characterized by a downstepping of peaks resembling a staircase pattern. However, the right-branching structures examined here were not uniform in how they were realized prosodically. In those structures in which the right-branch is not deeply embedded, as with noun phrases, two of three speakers produced one major phrase with an

additional boost on Peak 2. However, in the structures with the deeply embedded right branch, as in the relative clause constructions, all speakers produced two major phrases accompanied by a reset in pitch range. Given this apparent lack of consistency among the two constructions, in an account of the relationship between syntactic and prosodic structures, the depth of embedding is an important issue to consider. Another thing to consider is how we want to represent the influence of the syntax on prosodic structure. The theory of prosodic structure advocated by Ladd and Kubozono aims to encode syntactically sensitive variations in pitch register into the phonological representation, while alternative theories such as that of Beckman and Pierrehumbert wish to leave those decisions up to the pragmatics or discourse structure. A model that encodes the syntax necessarily complicates the phonological representation and cannot account for speaker variation. On the other hand, a model which leaves everything up to the pragmatics, without a concrete model of pragmatic or discourse structure, leaves too many degrees of freedom, and lessens the predictive capabilities.

It is clear that there is much room for future research in this area. An elaboration and more careful documentation of the phenomenon of metrical boost is necessary in order to examine its relation with syntactic branching structures, or more likely, pragmatic or discourse structures which are probably only vaguely related to syntactic constituency. Also, downstepping patterns in utterances with more diverse syntactic structures need to be thoroughly examined in order to assess the contribution of depth of embedding to the prosodic juncture between adjacent words. Lastly, a coherent model of discourse structure and its relation to the prosodic structure would prove extremely useful in determining the relative weights of contribution from the syntax and pragmatics to the overall prosodic realization of an utterance.

## References

- Azuma, J. & Tsukuma, Y. (1990) Prosodic features marking the major syntactic boundary of Japanese: A study on syntactically ambiguous sentences of the Kinki dialect. *Proceedings of the International Conference on Spoken Language Processing*, Kobe, 453-455.
- Azuma, J. & Tsukuma, Y. (1991) Role of F0 and pause in disambiguating syntactically ambiguous Japanese sentences. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 3: 274-277.
- Beckman, M.E. & Pierrehumbert, J.B. (1986) Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255-309.
- Cooper, W.E., Paccia, J.M. & Lapointe, S.G. (1978) Hierarchical Coding in Speech Timing. *Cognitive Psychology*, 10, 154-177.
- Hermes, D.J. & van Gestel, J.C. (1991) The frequency scale of speech intonation. *Journal of the Acoustical Society of America*, 90, 97-102.
- Jun, S.-A. (1993) *The phonetics and phonology of Korean prosody*. Ph.D. dissertation (Ohio State University).
- Klatt, D.H. (1975) Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 3, 129-140.
- Kori, S. (1992) Nihongo bun'onchoo no kenkyuu kadai. *Proceedings of the International Symposium on Prosody*. Nara, Japan.
- Kubozono, H. (1988) *The organization of Japanese prosody*. Ph.D. dissertation (University of Edinburgh).
- Kubozono, H. (1989) Syntactic and rhythmic effects on downstep in Japanese. *Phonology*, 6, 39-67.

- Kubozono, H. (1992) Modeling syntactic effects on downstep in Japanese. In *Papers in Laboratory Phonology II: Segment, Gesture, Tone* (G.J. Docherty & D.R. Ladd, editors), pp. 368-387. Cambridge: Cambridge University Press.
- Ladd, D.R. (1986) Intonational phrasing: The case for recursive prosodic structure. *Phonology Yearbook*, 3, 311-340.
- Ladd, D.R. (1988) Declination "reset" and the hierarchical organization of utterances. *Journal of the Acoustical Society of America*, 84, 530-544.
- Ladd, D.R. (1990) Metrical representation of pitch register. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (J. Kingston & M.E. Beckman, editors), pp. 35-57. Cambridge: Cambridge University Press.
- Ladd, D.R. (1993) In defense of a metrical theory of downstep. In *The Phonology of Tone: The Representation of Tonal Register* (H. van der Hulst & K. Snider, editors), pp. 109-132. Mouton de Gruyter.
- Ladd, D.R. & Johnson, C. (1987) 'Metrical' factors in the scaling of sentence-initial accent peaks. *Phonetica*, 44, 238-245.
- Lehiste, I. (1973) Phonetic disambiguation of syntactic ambiguity. *Glossa*, 7, 107-122.
- Lehiste, I., Olive, J.P. & Streeter, L.A. (1976) The role of duration in disambiguating syntactically ambiguous utterances. *Journal of the Acoustical Society of America*, 60, 1199-1202.
- Maekawa, K. (1991) Perception of intonational characteristics of WH and non-WH question in Tokyo Japanese. *Proceedings of the XIIth International Congress of Phonetic Sciences*, 4/5: 202-205.
- McCawley, J. (1968) *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton.
- Moore, B.C.J. & Glasberg, B.R. (1983) Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74, 750-753.
- Nespor, M. and I. Vogel (1986) *Prosodic Phonology*. Dordrecht: Foris Publications.
- Pierrehumbert, J.B. & Beckman, M.E. (1988) *Japanese Tone Structure*. Cambridge, Massachusetts: MIT Press.
- Poser, W. (1984) *The Phonetics and Phonology of Tone and Intonation in Japanese*. Ph.D. dissertation (Massachusetts Institute of Technology).
- Poser, W. (1990) Word-internal phrase boundary in Japanese. In *The Phonology-Syntax Connection* (S. Inkelas & D. Zec, editors), pp. 279-287. Chicago: University of Chicago Press.
- Selkirk, E. (1984) *Phonology and Syntax: The Relation between Sounds and Structure*. Cambridge, Massachusetts: MIT Press.
- Selkirk, E. (1986) On Derived Domains in Sentence Phonology. *Phonology*, 3, 371-405.
- Selkirk, E. & Tateishi, K. (1988) Constraints on minor phrase formation in Japanese. *Proceedings of the Chicago Linguistic Society*, 24, 316-336.
- Selkirk, E. & Tateishi, K. (1991) Syntax and downstep in Japanese. In *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda* (C. Georgopoulos & R. Ishihara, editors), pp. 519-543. Dordrecht: Kluwer Academic Publishers.
- Streeter, L.A. (1978) Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, 64, 1582-1592.
- Terken, J. & Collier, R. (1992) Syntactic influences on prosody. In *Speech Perception, Production and Linguistic Structure* (Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka, editors), pp. 427-438. Tokyo: IOS Press.

- Uyeno, T., Hayashibe, H. & Imai, K. (1979) On pitch contours of declarative, complex sentences in Japanese. *Research Institute of Logopedics and Phoniatrics*, 13, 175-187. University of Tokyo.
- Uyeno, T., Hayashibe, H., Imai, K., Imagawa, H. & Kiritani, S. (1980) Comprehension of relative clause construction and pitch contours in Japanese. *Research Institute of Logopedics and Phoniatrics*, 14, 225-236. University of Tokyo.
- Uyeno, T., Hayashibe, H., Imai, K., Imagawa, H. & Kiritani, S. (1981) Syntactic structures and prosody in Japanese: A study on pitch contours and the pauses at phrase boundaries. *Research Institute of Logopedics and Phoniatrics*, 15, 91-108. University of Tokyo.
- Venditti, J.J. & Yamashita, H. (in preparation) The prosodic characteristics of temporarily ambiguous constructions in Japanese. ms. Ohio State University.

