DOCUMENT RESUME

ED 409 733 FL 024 673

AUTHOR	Ainsworth-Darnell, Kim, Ed.; D'Imperio, Mariapaola, Ed.
TITLE	Papers from the Linguistics Laboratory. Working Papers in Linguistics, No. 50.
INSTITUTION	Ohio State Univ., Columbus. Dept. of Linguistics.
PUB DATE	Jul 97
NOTE	183p.
PUB TYPE	Collected Works - General (020)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	Articulation (Speech); Consonants; Contrastive Linguistics; Diachronic Linguistics; English; Individual Differences; Italian; Japanese; *Language Patterns; *Language Processing; Language Research; Language Styles; Language Variation; Linguistic Theory; Mass Media; North American English; *Oral Language; *Phonology; Radio; Russian; Suprasegmentals; Syntax; Uncommonly Taught Languages
IDENTIFIERS	*Balinese; Diphthongs; Ohio (Columbus)

ABSTRACT

Research reports included in this volume of working papers in linguistics are: "Perception of Consonant Clusters and Variable Gap Time" (Mike Cahill); "Near-Merger in Russian Palatalization" (Erin Diehm, Keith Johnson); "Breadth of Focus, Modality, and Prominence Perception in Neapolitan Italian" (Mariapaola D'Imperio); "The Northern Cities Shift in the Heartland? A Study of Radio Speech in Columbus, Ohio" (Steve Hartman Keiser, Frans Hinskens, Bettina Migge, Elizabeth A. Strand); "Syntactically-Governed Accentuation in Balinese" (Rebecca Herman); "The Auditory-Perceptual Basis for Speech Segmentation" (Keith Johnson); Production and Perception of Individual Speaking Styles" (Keith Johnson, Mary E. Beckman); "Japanese ToBI Labelling Guidelines" (Jennifer J. Venditti); and "A Cross-Linguistic Study of Diphthongs in Spoken Word Processing in Japanese and English" (Kiyoko Yoneyama). Individual papers contain references. (MSE)

Working Papers in Linguistics

No. 50

PAPERS FROM THE LINGUISTICS Laboratory

Edited by

Kim Ainsworth-Darnell
Mariapaola D'Imperio

July 1997

The Ohio State University
Department of Linguistics

The Ohio State University

Working Papers in Linguistics No. 50

Papers from the Linguistics Laboratory

Edited by

Kim Ainsworth-Darnell
Mariapaola D'Imperio

The Ohio State University

Department of Linguistics

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

July 1997

Information Concerning the OSUWPL

The Working Papers in Linguistics is an occasional publication of the Department of Linguistics of the Ohio State University and usually contains articles written by students and faculty of the department. There are generally one to three issues per year. Information about available issues appears below. Numbers 1, 5, 10 and 23 are out of print and no longer available. The tables of contents can also be viewed on our World Wide Web server:

http://ling.ohio-state.edu/Department/Osuwpl.html

There are two ways to subscribe to OSUWPL. The first is on a regular basis: the subscriber is automatically sent and billed for each issue as it appears. The second is on an issue-by-issue basis: the subscriber is notified in advance of the contents of each issue, and returns an order blank if that particular issue is desired.

Volume

18 $5.00. 183 pp. (June 1975): Papers by Michael Geis, Sheila Geoghegan, Jeanette Gundel, G.K. Pullum, and Arnold Zwicky.

19 $5.00. 214 pp. (September 1975), edited by Robert K. Herbert: *Patterns of Language, Culture and Society: Sub-Saharan Africa* contains eighteen papers presented at the Symposium on African Language, Culture, and Society, held at The Ohio State University on April 11, 1975.

20 $5.00. 298 pp. (September 1975), edited by Robert K. Herbert: *Proceedings of the Sixth Conference on African Linguistics* contains twenty-seven papers presented at the Sixth Conference on African Linguistics, held at The Ohio State University on April 12-13, 1975.

21 $5.00. 252 pp. (May 1976), edited by Arnold M. Zwicky: *Papers on Nonphonology*. Papers by Steven Boer and William Lycan, Marian Johnson, Robert Kantor, Patricia Lee, and Jay Pollack.

22 $5.00. 151 pp. (February 1977), edited by Olga Garnia: *Papers in Psycholinguistics and Sociolinguistics*. Papers by Sara Garnes, Olga Garnica, Mary Louise Edwards, Roy Major, and John Perkins.

24 $5.00. 173 pp. (March 1980), edited by Arnold M. Zwicky: *Clitics and Ellipsis*. Papers by Robert Jeffers, Nancy Levin (OSU Ph.D. Dissertation), and Arnold Zwicky.

25 $5.00. 173 pp. (January 1981), edited by Arnold M. Zwicky: *Papers in Phonology*. Papers by Donald Churma, Roderick Goman (OSU Ph.D. Dissertation), and Lawrence Schourup.

26 $5.00. 133 pp. (May 1982), edited by Brian D. Joseph: *Grammatical Relations and Relational Grammar*. Papers by David Dowty, Catherine Jolley, Brian Joseph, John Nerbonne, and Amy Zaharlick.

27 $5.00. 164 pp. (May 1983), edited by Gregory T. Stump: *Papers in Historical Linguistics*. Papers by Donald Churma, G.M. Green, Leena Hazelkorn, Gregory Stump, and Rex Wallace.

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, Sook-hyang 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 Theor* 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.

45 $15.00. 223 pp. (February 1995), edited by Stefanie Jannedy: *Papers from the Linguistics Laboratory.* Papers by Julie E. Boland & Anne Cutler, K. Bretonnel Cohen, Rebecca Herman, Stefanie Jannedy, Keith Johnson & Mira Oh, Hyeon-Seok Kang, Jaan Ross & Ilse Lehiste, Ho-Hsien Pan, and Shu-hui Peng.

46 $12.00. 114 pp. (Spring 1995), edited by Elizabeth Hume, Robert Levine, and Halyna Sydorenko: *Studies in Synchronic and Diachronic Variation.* Papers by Mary Bradshaw, Brian Joseph, Hyeree Kim, Bettina Migge, and Halyna Sydorenko.

47 $12.00. 134 pp. (Autumn 1995), edited by David Dowty, Rebecca Herman, Elizabeth Hume, and Panayiotis A. Pappas. *Varia.* Papers by Kim Ainsworth-Darnell, Qian Gao, Karin Golde, No-Ju Kim, David Odden, and Arnold Zwicky.

48 $15.00. 227 pp. (Spring 1996), edited by David Dowty, Rebecca Herman, Elizabeth Hume, and Panayiotis A. Pappas. Papers in Phonology. Papers by Mary Bradshaw, Mike Cahill, Rebecca Herman, Hyeon-Seok Kang, Nasiombe Mutonyi, David Odden, Frederick Parkinson, Robert Poletto, and R. Ruth Roberts-Kohno.

49 $15.00. 177 pp. (Summer 1996), edited by Jae-Hak Yoon and Andreas Kathol. *Papers in Semantics.* Papers by Mike Calcagno, Chan Chung, Alicia Cipria and Craige Roberts, Andreas Kathol, Craige Roberts, Eun Jung Yoo, and Jae-Hak Yoon.

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)

Information Concerning OSDL
(Ohio State Dissertations in Lingnistics)

Ohio State Linguistics Students are now making available dissertations written by our members since 1992. For more information regarding available titles and abstracts as of July 1997, please visit our World Wide Web server at

http://ling.ohio-state.edu/Department/Dissertations.html

For more information and ordering procedures, please contact osdl@ling.ohio-state.edu or

OSDL
Department of Linguistics
The Ohio State University
222 Oxley Hall
1712 Neil Avenue
Columbus, Ohio 43210-1289
U.S.A.

Currently available titles (July 1997):

Benjamin Xiaoping Ao (1993). *Phonetics and Phonology of Nantong Chinese.* (Advisor: David Odden)

Gayle Ayers (1996). *Nuclear Accent Types and Prominence: Some Psycholinguistic Experiments.* (Advisor: Mary E. Beckman)

Hee-Rahk Chae (1992). *Lexically Triggered Unbounded Discontinuities in English: An Indexed Phrase Structure Grammar Approach.* (Advisor: Arnold Zwicky)

Chan Chung (1995). *A Lexical Approach to Word Order Variation in Korean.* (Advisor: Carl Pollard)

John Xiang-ling Dai (1992). *Chinese Morphology and its Interface with the Syntax.* (Advisor: Arnold Zwicky)

Hyeon-Seok Kang (1997). *Phonological Variation in Glides and Diphthongs of Seoul Korean: Its Synchrony and Diachrony.* (Advisor: Donald Winford)

Sun-Ah Jun (1993). *The Phonetics and Phonology of Korean Prosody.* (Advisor: Mary E. Beckman)

Hyeree Kim (1996). *The Synchrony and Diachrony of English Impersonal Verbs: A Study in Syntactic and Lexical Change.* (Advisor: Brian D. Joseph)

No-Ju Kim (1996). *Tone, Segments, and Their Interaction in North Kyungsang Korean: A Correspondence Theoretic Account*. (Advisor: David Odden)

Gina Maureen Lee (1993). *Comparative, Diachronic and Experimental Perspectives on the Interaction between Tone andd Vowel in Standard Cantonese*. (Advisor: Brian D. Joseph)

Sook-hyang Lee (1994). *A Cross-Linguistic Study of the Role of the Jaw in Consonant Articulation*. (Advisor: Mary E. Beckman)

Frederick B. Parkinson (1996). *The Representation of Vowel Height in Phonology*. (Advisor: David Odden)

Shu-Hui Peng (1996). *Phonetic Implementation and Perception of Place Coarticulation and Tone Sandhi*. (Advisor: Mary E. Beckman)

Katherine Welker (1994). *Plans in the Common Ground: Toward a Generative Account of Conversational Implicature*. (Advisor: Craige Roberts)

Jae-Hak Yoon (1996). *Temporal Adverbials and Aktionsarten in Korean* (downloadable). (Advisor: Craige Roberts)

Foreword

This volume includes papers that take an experimental or quantitative approach to linguistics. It is the sixth in a series of progress reports from the OSU Linguistics Laboratory (see also OSUWPL No. 36, 38, 43, 44, and 45). Some of the papers are in progress or have been presented at international conferences, while others have been submitted for publication to professional journals. We would like to thank Keith Johnson and Rebecca Herman for their help during the compilation of this volume. The production of this volume was supported by the National Institute on Deafness and other Communicatios Disorders under Grant No. R29DC01645-05 and by The Ohio State University Department of Linguistics.

Kim Ainsworth-Darnell
Mariapaola D'Imperio
July 1997

Researchers in Phonetics... in the OSU Linguistics Laboratory

Kim Ainsworth-Darnell
Mary E. Beckman
Jose Benki
Julie E. Boland
Kevin B. Cohen
Mariapaola D'Imperio
Donna Erickson
Steve Hartman Keiser
Rebecca Herman
Stefanie Jannedy
Keith Johnson
Ilse Lehiste
Matt Makashay

Bettina Migge
Joyce McDonough
Norma Mendoza-Denton
Fox Mulder
Amanda Miller-Ockhuizen
Panaiyotis A. Pappas
R. Ruth Roberts-Kohno
Elizabeth A. Strand
Jennifer J. Venditti
Stephen Winters
Kiyoko Yoneyama
Teruhisa Uchida
Takashi Otake

Ohio State University Working Papers in Lingnistics No. 50

Papers from the Linguistics Laboratory

Table of Contents

OSU Working Papers in Linguistics 50 (1-10)

Perception of Consonant Clusters and Variable Gap Time*

Mike Cahill
cahill@ling.ohio-state.edu

Abstract: In every case in which measurements of labial-velar stops [kp, gb] have been made, it has been found that the labial and velar gestures are not strictly simultaneous, but rather that the velar gesture slightly precedes the labial one (thus [k͡p] and not [p͡k]). One possible explanation for this is that [k͡p] is more perceptually salient than [p͡k]. This paper reports an attempt to test this hypothesis by observing listeners' identifications of [apka] and [akpa] with variable gap times inserted between the consonantal onset and release. The results showed that [apka] was more readily identified than [akpa], effectively showing that perceptual salience cannot be invoked to explain the ordering of velar and labial gestures in labial-velar stops.

INTRODUCTION

Labial-velar stops [kp], [gb] occur in many languages from central and west Africa, where the bulk of them are found, as well as in a handful of languages in and around Papua New Guinea.[1] They are commonly described as having "simultaneous" closure at the labial and velar places of articulation. However, most transcriptions have recorded them as [k͡p] and not [p͡k], and this is no accident. Spectrographic evidence shows that a vowel preceding a labial-velar stop makes a transition into a velar component, and the release of the consonant has labial characteristics, in languages as diverse as Dedua (from Papua New Guinea) and Efik (Ladefoged & Maddieson 1996) and Ibibio (Connell 1994), both from west Africa. Also, Maddieson has presented direct evidence that in Ewe, at least, the labial gesture both starts and ends later than the velar gesture, as in Figure 1 below, taken from electromagnetic articulography data in Maddieson (1993) .

* My very great thanks to Keith Johnson, who provided invaluable aid in the design and implementation of every stage of the experiment reported herein.

[1] A few Creole languages of the Caribbean, such as Ndyuka of Surinam, also have labial-velar stops, presumably as a result of African language substrata (Huttar & Huttar 1994).

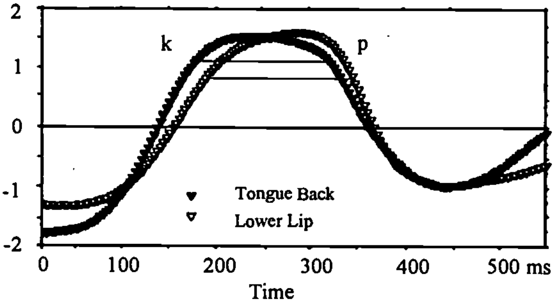


Fig. 1 Coordination of lower lip and tongue back movements in the Ewe word **akpa**. Y-axis is vertical displacement; horizontal lines indicate the likely duration of actual contact of the articulator. (Maddieson 1993)

In this paper, then, [k͡p] and [p͡k] will both refer to articulations that are mostly overlapping, but in [k͡p] the labial gesture follows the velar one, as above, while in [p͡k], the labial gesture would precede the velar one. I will focus on the voiceless stop [k͡p] here, though the discussion is also applicable to the voiced labial-velar stop [g͡b].

One of the questions arising out of research on labial-velar stops is why this partial or incomplete overlap should exist, rather than total simultaneity. Also, why should it be that universally (as far as we know) there is an asymmetry of gestural overlap and that this asymmetry should always be in the same direction?

At least three possibilities exist and are worthy of consideration. One possibility is that there is an "ease of articulation" factor, that is, that [k͡p] requires less effort to produce than [p͡k]. One might argue that with the condyles of the jaw acting as a pivot, especially if a consonant is pronounced after a vowel, it would be more natural for the articulators closer to the pivot point to make contact sooner than those further away; so a consonant made with a place of articulation further back in the mouth would be more likely to precede a consonant made in the front of the mouth. This could be a physiological explanation of the data in Hume (1996), who gives examples from several languages in which metathesis of consonant clusters operates to give an output in which the more posterior consonant precedes the more anterior one. She proposes a phonological constraint in which the more posterior of a consonant cluster pair is favored to precede the other. Arguing from the "ease of articulation" viewpoint, however, is notoriously suspect, since languages of the world abound in sounds which are not simple to produce. Sounds such as implosives, clicks, ejectives, and complex consonant clusters come to mind. A perusal of Ladefoged & Maddieson (1996) yields an abundance of examples. One may be able to argue persuasively and theoretically that certain sounds are, in fact, more difficult to make than others, but the existence of difficult sounds in languages of the the world makes this argument a tendency rather than a robust explanation for the phenomenon. Also, what is judged as "difficult" largely depends on the inventory of sounds in the speaker's native language compared to the language under consideration. I judge the [kp] found in most Ghanaian languages to be fairly difficult, while the Ghanaian takes the [k͡p] in stride but judges the phonetics of American English 'squirrel' to be very difficult.

Another possibility is that since the historical development of labial-velars seems commonly, perhaps always, to be the reflex of a labialized stop, this labial release has been maintained in modern languages. The two main sources of /kp/ historically seem to be sometimes ***pw**, exemplified by Aghem (Hyman 1979), but more often ***kw**, exemplified by the Sawabantu group of languages in western Cameroon, (Mutaka and Ebobissé 1996). These and other examples are examined in Cahill (in prep). In both of these proto-forms, the release is labial, whereas the start of the consonant, at least in the case of ***kw**, is not. The possibility is that the asymmetry present in the proto-form, that is, the labiality being skewed more to the release, is preserved in the synchronic reflexes.

A third possibility is that, perceptually, a [k͡p] is more salient than a [p͡k]. This implies that [k͡p] is easier to perceive in some way than a [p͡k].[2] Again, Hume's constraint favoring consonant clusters with the more posterior consonant occurring first would be a way of expressing this tendency in phonological terms. But, as with the "ease of articulation" possibility, languages abound which have hard-to-hear sounds. These would include creaky and breathy vowels, different fricatives, labial-velars themselves, and a host of others. Similarly to the point made for difficult articulations, sounds which are judged "hard to perceive" are mainly those which do not occur in the hearer's native language.

It is of course possible that more than one of the above factors could be at work here. For example, [k͡p] could have developed for historical reasons, then remained as such for ease of articulation. Each possibility also has its own set of objections. In addition, it is hypothetically possible that [k͡p] exists rather than [p͡k] merely as an accident of language, though the universality of [k͡p] makes this scenario rather dubious. But if we assume that there is a reason (or reasons) behind the asymmetry of labial-velars, then it should be possible to investigate what that reason is, despite any initial difficulties.

The experiment reported here was an attempt to test the third hypothesis. More specifically, the hypothesis this experiment addressed was that the reason why the partial overlap of labial-velars is always skewed in the direction of labial release is that a labial release and velar onset is more perceptually salient than a velar release and labial onset. To test this, we spliced together sequences of [kp] and [pk], with varying gap durations, and tested to see which was more readily identifiable.

The result of the experiment did not support this hypothesis, but showed rather that [pk] was the more salient of the two clusters.

[2] Chomsky and Halle (1968), in reasoning about perception of multiply-articulated sounds, get the phonetics precisely backwards with respect to labial-velar stops. They write, "The order of release of the different closures is governed by a simple rule. In sounds without supplementary motions [i.e. movement of the glottis during the period of closure- mc], the releases are simultaneous. In sounds produced with supplementary motions, closures are released in the order of increasing distance from the lips. The reason for this ordering is that only in this manner will clear auditory effects be produced, for acoustic effects produced inside the vocal tract will be effectively suppressed if the vocal tract is closed." (1968:324). This predicts labial-velars with a simple pulmonic airstream should release both closures simultaneously, while labial-velars with an ingressive velaric airstream should release the labial closure first. However, in both cases, it is the labial closure which is released last (see Ladefoged 1968, Painter 1970).

METHOD

Spectrographic studies of labial-velar stops have shown a release burst characteristic of labial stops, but a transition from the preceding vowel characteristic of velar stops (Ladefoged 1968, Garnes 1975, Connell 1991, 1994, Ladefoged & Maddieson 1996). So splicing together the consonants [k] and [p] gives a reasonable facsimile of a labial-velar stop.

To produce the test sounds, the author recorded several tokens of the syllables [ap], [ak], [ka], and [pa] in a soundproof recording booth using a Marantz 220 tape recorder, with a Sure SM-48 microphone. Representatives of the appropriate tokens were then spliced together using the CSpeech program to form the output tokens [ap-pa], [ap-ka], [ak-ka], and [ak-pa]. (These stimuli will be referred to below with capitals, e.g. APKA.) Silent durations of 0-200 ms, in 25 ms increments, were inserted between the offset of voicing in [aC] and the release burst of [Ca]. For the [ap-ka] and [ak-pa] tokens, intervals were extended to 400 ms as well. A VisualBasic program was set up, so that listeners heard the tokens in random order over headphones, and selected either *apa, aka, akpa,* or *apka* as the closest to what they heard.

15 listeners participated in the experiment, all undergraduate students from Ohio State University. All but one had English as their mother tongue and were from Ohio. The exception was a Jordanian student whose first language was Arabic, but her responses were not markedly different from the others, so they are included as well.

The listeners were seated in a soundproof booth, with a computer screen in front of them. The program played the token, and the subject used the computer mouse to click on a button on the screen labeled **apa**, **aka**, **akpa**, or **apka**. There was a 2-second interval after they clicked before the next token was played. The 47 tokens were randomized; when one block of 47 trials finished, another block began. The same set of tokens was repeated in this way four times, with a different randomized order each time, for a total of 168 total tokens presented to each subject. The experiment was self-paced, with a token not presented until a response was given to the previous one. The total time for each run of the experiment ranged from 25-40 minutes.

RESULTS

Figure 2 shows the two most common responses to the AKPA stimuli. The main trend is that at shorter gap durations, the AKPA stimulus was perceived as "apa." As the gap duration increased, the perception of "akpa" also increased. The two responses were approximately equal at about 80ms, as measured by the crossover point of the two plotted curves. As expected, virtually all the non-"akpa" responses were "apa", having the same release as the stimulus; therefore, the very few responses which were "aka" or "apka" are not plotted.

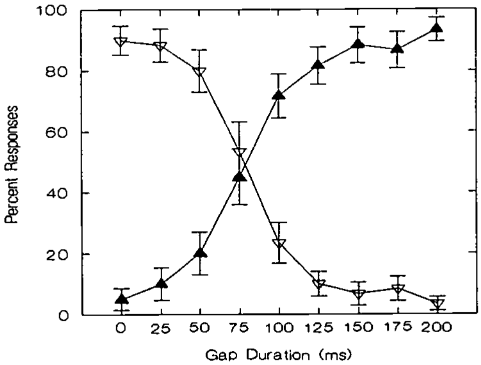


Fig. 2 Responses to AKPA input: filled = "akpa", open = "apa"

Figure 3 shows the responses to the APKA stimuli. Similar to the above, for shorter intervals, the "aka" response was more often given; as the gap interval increased, the "akpa" response was increasingly given. As expected, virtually all the non-"apka" responses were "aka", having the same release as the stimulus; therefore, the very few responses which were "apa" or "akpa" are not plotted in Fig. 3. In comparison to the AKPA stimulus, the APKA stimulus was correctly identified at a shorter gap duration; the crossover from "aka" to "apka" occurred at only 25 ms (compared to 80 ms for AKPA).

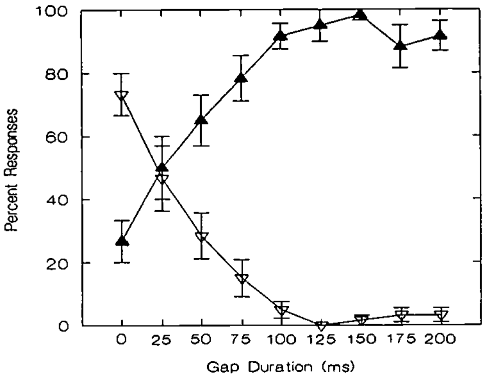


Fig. 3 Responses to APKA input: filled = "apka", open = "aka"

Figure 4 contains the "akpa" and "apka" curves from Figures 2 and 3, showing directly that the "apka" response to APKA was chosen at shorter gap durations than the "akpa" response to AKPA.

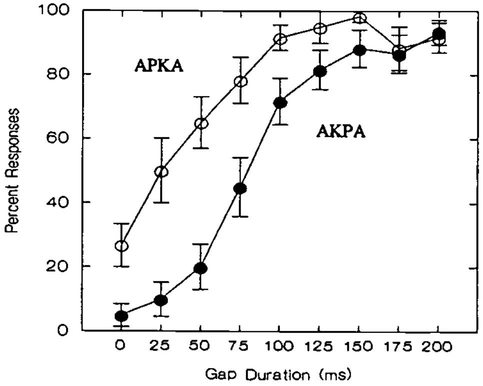


Fig. 4 Responses to KP vs. PK stimuli: filled = AKPA/akpa, open = APKA/apka

A result which was unexpected was that for both the AKKA and APPA stimuli, at very long intervals, subjects occasionally identified the stimulus as a heterogeneous cluster "apka" or "akpa." This is shown in Figs. 5-6. As above, the release consonant of the response was the same as the stimulus for almost all responses. The bulk of the mis-responses was from three subjects, but there was some scattered similar response from others as well.

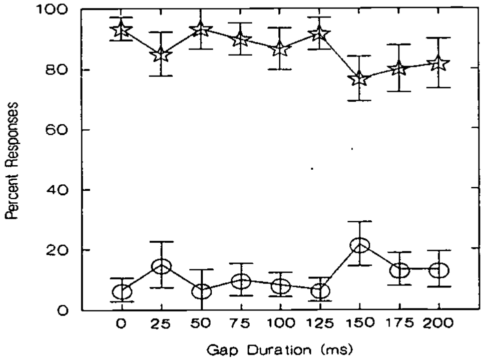


Fig. 5 Responses to AKA stimuli: star = "aka", circle = "apka"

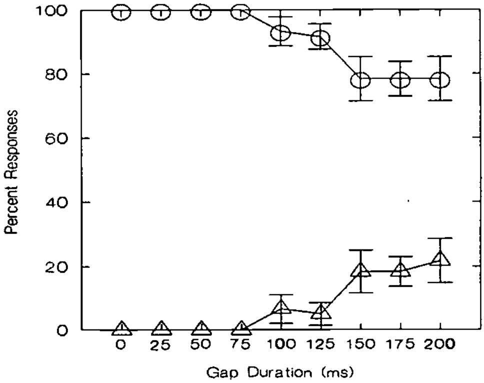


Fig. 6 Responses to APA stimuli: circle = "apa", triangle = "akpa"

DISCUSSION

Several conclusions may be drawn from this data.

First, at long gap durations the subjects sometimes heard a phonetic geminate as a cluster of heterogeneous consonants (Figs. 5-6). In English, geminate stop consonants are rare (the *k:* in *bookkeeper* being one example), and do not contrast with non-geminates in non-compound words at all. Heterogeneous clusters, on the other hand, are more common. The hearers, in common with any speakers of one particular language hearing unfamiliar sounds in a second language, evidently tried to "fit" the unfamiliar sound into their native sound system.

Second, since in the short gap durations [kp] was identified as [p], and [pk] was identified as [k], the release is shown to be more crucial than the onset in identifying the nature of the consonants in question (Figs. 2-4). This is consistent with much other research in this area (for a summary, see Pickett, Bunnell, & Revoile 1995 and references cited therein).

Besides the immediate question at hand, this tendency of the release being the key to identifying a consonant helps explain a common diachronic tendency. Labial-velars historically tend to change to simple labials, e.g. *kp > p, *gb > b (Cahill in prep and references therein). This can be explained on the perceptual grounds that labial-velars have labial releases and thus tend to be readily perceived as labials.[3]

Third, [kp] was mis-identified at longer gap durations than [pk], or to state it in positive terms, the English speakers in this experiment identified the [pk] cluster more readily than they did [kp] (Figs. 2-4). To what can we attribute this difference? One possibility is that the transition from the vowel into the consonant is the key factor in identifying the consonant, and that [ap] was more salient than [ak]. However, we saw above that the release is more crucial in identifying the consonant. If the release is indeed more crucial, then a reasonable conclusion is that the [p] release of [kp] was less perceptible than the [k] release of [pk].[4] This conclusion effectively falsifies the beginning hypothesis. Recall that the hypothesis, looking for an explanation of why labial-velar consonants in the world's languages are seemingly universally [k͡p] and not [p͡k], postulated that a [kp] would be more identifiable than the reverse [pk]. Our results show exactly the opposite: [pk] is more identifiable than [kp].

Some potential complicating factors exist in this experiment and must be addressed. These include possible interference from English phonotactic statistics, possible asymmetry of the recorded stimuli, and how close the experiment was to the phonetics of real-language [k͡p].

The first factor has to do with the relative frequencies of [kp] vs. [pk] in English words. If [pk] is significantly more frequent than [kp], we might expect listeners to choose the more familiar [pk] and thus skew the results in favor of that selection. However, a search of the MRC database of approximately 150,000 English words showed 3 with [kp] and 7 with [pk] between vowels. This is a difference, to be sure, but not enough to account for the patterns found in this experiment.[5]

[3] The same point is made in Connell (1991, 1994).

[4] However, in Figs. 4-5, [p] was never mis-identified at short gap intervals, while [k] was. So do Figs. 4-5 show that [p] is more perceptible? I do not believe so, because of the very small numbers involved in the mis-identification.

[5] The words were: [kp] - jackpot, pickpocket, stockpile, [pk] - napkin, pipkin, shopkeeper, tipcat, upcast, upcountry, upkeep. I can add "backpack" and perhaps others to the [kp] list, but the fact remains that both [VkpV] and [VpkV] are relatively rare patterns in English, and occur almost exclusively with compound words. A further check shows that this pattern of relatively rare, mostly compound, words holds for most other

Another question to consider is were the recorded segments [pa] and [ka] essentially equal in peripheral properties that could affect perceptibility, or not? For example, in the spectrograms of [apka] and [akpa] (Fig. 7), the aspiration of [ka] is longer than that of [pa]. This probably does not contribute to the higher salience of [ka] in this experiment, since both cross-linguistically and specifically in English, a [k] followed by vowel has a significantly longer VOT than a [p] or [t] (Lisker & Abramson 1964). Thus the stimuli for this experiment fit within normal ranges. Other factors, such as loudness, also do not appear relevant.

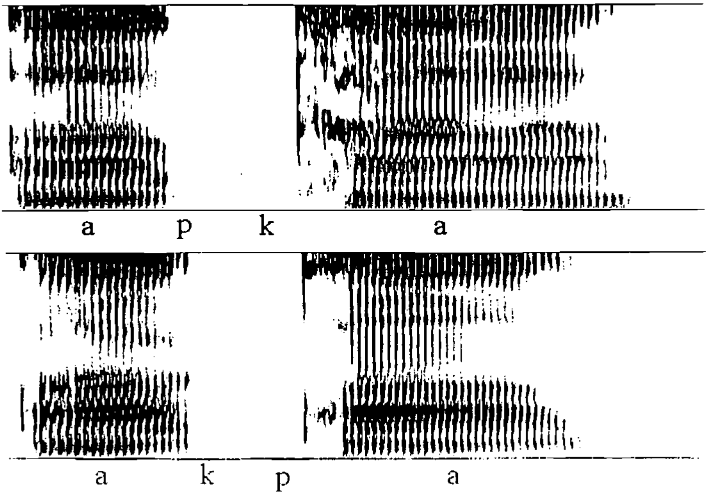


Fig. 7 Spectrograms of the [apka] and [akpa] stimuli

A further question is whether the vowel quality had any impact on the results. Would the same results be found with other vowels besides [a]?

A more serious concern is how similar the spliced [kp] in this experiment is to a real-language [k͡p]. There are at least two differences between the spliced [kp] here and at least some [k͡p] in natural languages. Quite often, literature sources note that [kp] is unaspirated, unlike simple voiceless stops (e.g. Smith 1967). The input sounds for this experiment had normal English aspiration (see Fig. 7). Also, there is often a suction mechanism for producing labial-velars (Ladefoged 1968), which was absent in this experiment. These factors could conceivably have an impact on the results.

stop-stop sequences, in whatever combination of voiced or voiceless, with the exception of [kt] and [pt], which are relatively more common.

CONCLUSION

In this experiment, [pk] was shown to be more perceptually salient than [kp]. If this result is a universal of human language perception, this conflicts with the universal of language production that in languages with labial-velars, [k͡p] is preferred over [p͡k]. In the introduction, we discussed three possible factors which could account for the universality of [k͡p]: first, perceptual salience, which was tested in this experiment; second, a relic of labial-velars' historical development from labialized stops (e.g. k^w); and third, greater ease of articulation for [k͡p] than [p͡k]. We have shown here that the bias of perceptual salience is actually towards [pk], not the reverse. We conclude, then, that the perceptual salience of a velar release is not a factor in the universality of [k͡p], but must be overridden by the historical or articulatory factors previously discussed. Future experiments and studies are needed to focus on these influences.

REFERENCES

Cahill, Mike. (in preparation) Labial-velars: Phonetics, History, and Phonology. ms., OSU.

Connell, Bruce. (1991). Accounting for the reflexes of labial-velar stops. In Rossi, M. (ed.) Proceedings of the XIIth ICPhS, Aix-en-Provence, Vol. 3:110-113.

Connell, Bruce. (1994). The structure of labial-velar stops. Journal of Phonetics 22:441-476.

Garnes, Sara. (1975). An Acoustic Analysis of Double Articulations in Ibibio. Proceedings of the Sixth Conference on African Linguistics, OSU Working Papers in Linguistics No. 20.

Hume, Elizabeth. (1996) A Non Linearity Based Account of Leti Metathesis. ms, OSU.

Huttar, George L., and Mary L. Huttar. (1994). Ndyuka. London: Routledge.

Hyman, Larry M. (1979). Phonology and noun structure. In Hyman, Larry M. (Ed) Aghem Grammatical Stucture. Southern California Occasional Papers in Linguistics No. 7. USC.

Ladefoged, Peter. (1968). A Phonetic Study of West African Languages. (2nd ed.) Cambridge University Press.

Ladefoged, Peter, and Ian Maddieson. (1996). The Sounds of the World's Languages. Cambridge, MA: Blackwell Publishers.

Lisker, Leigh, and Arthur S. Abramson. (1964). A cross language study of voicing in initial stops: Acoustical measurements. Word 20:384-422.

Maddieson, Ian. 1993. Investigating Ewe articulations with electromagnetic articulography. Forshungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München 31:181-214.

Mutaka, Ngessimo M., and Carl Ebobissé. (1996) The formation of labio-velars in Sawabantu: evidence for feature geometry. ms., to appear in Journal of West African Languages.

Painter, Colin. (1970). Gonja: A Phonological and Grammatical Study. Bloomington: Indiana University.

Pickett, J.M., H. Timothy Bunnell, and Sally G. Revoile. 1995. Phonetics of Intervocalic Consonant Perception: Retrospect and Prospect. Phonetica 52:1-40.

Smith, N.V. (1967). The Phonology of Nupe. Journal of African Languages 6.2: 153-169

OSU Working Papers in Linguistics 50 (11-18)

Near-Merger in Russian Palatalization[*]

Erin Diehm
Keith Johnson
diehm.1@osu.edu, kjohnson@ling.ohio-state.edu

Abstract: This study investigates the palatalized consonants of Russian in environments which prove difficult for second language learners of Russian. To this end, we conducted a production and a perception study. In the production experiment, native and nonnative speakers demonstrated different patterns of contrast. Results of the perception experiment are surprising because the nonnative speakers were able to distinguish more phonetic contrasts than native speakers. The native-speakers' performance provides supportive evidence of a 'near merger', where a contrast is maintained in production but lost in perception.

INTRODUCTION

In Russian, palatalized consonants contrast with plain, unpalatalized consonants. (Avanesov, 1972; Bolla, 1981; Zinder, *et al.*, 1964; Panov, 1964) The Russian palatalized consonants, however, also occur in sequences which American students of Russian find difficult to distinguish from bare palatalization. The sequences that we investigated contrast bare palatalized consonants with palatalized consonants followed by the palatal glide, and with palatalized consonants followed by the high front vowel followed by the palatal glide. We will call these the 'palatalized' (C^jV), the 'palatalized-plus-jot' (C^jjV) and the 'palatalized-i-jot' (C^jijV), respectively.

Of particular interest is the fact (which we will assume for the time being) that the contrast between palatalized-plus-jot sequences and palatalized-i-jot sequences, when stressed word-finally, is in a state of near-merger in Russian.[1] This contrast is rarely used to distinguish words and native-speakers intuitively feel that there is no real difference between them. We will present experimental evidence supporting this assumption. We will also argue that the acquisition of near-mergers (or at least this one) provides evidence that native-speakers and second language learners adopt very different perceptual strategies.

To anticipate our results, in an acoustic-phonetic production study of Russian plain (CV) and palatalized (C^jV, C^jjV, and C^jijV) consonants (totaling four sequence types) we found that learners did not distinguish all of the sequence types that native speakers do. However, in a perception study we found that native speakers failed to distinguish the palatalized-plus-jot and the palatalized-i-jot sequences (even though they did pronounce them differently) while our group of learners did attend to this distinction.

We will suggest that the native-speakers' behavior in the perception portion of the experiment reflects knowledge of the relative functional weight of the distinctions being

[*] Presented at the 1997 Meeting of the Linguistic Society of America, January 5, 1997.

[1] We are using the term 'near-merger' in a relativey non-traditional way to refer to sequences which are spelled differently, pronounced slightly differently and judged by native speakers to not contrast.

investigated - that is, native-speakers do not treat a near-merger with the same attention that they reserve for normal contrasts. Thus, the native speakers listened 'linguistically' while the learners tended to listen 'phonetically'.

ACOUSTIC STUDY

METHODS

Thirty speakers participated in the production study. Sixteen (8 women, 8 men) were native speakers of Russian, and fourteen (8 women, 6 men) were American learners of Russian. These learners had studied Russian for at least 5 years and had lived in Russia for at least 4 months.

The speakers read a word list that was composed of near-minimal sets such as those illustrated in Table 1. Each set illustrated contrasts between a plain consonant, its palatalized counterpart, the palatalized consonant followed by the palatal glide, and the palatalized consonant followed by the high front vowel [i] and then the palatal glide. The word list contained 14 examples of each of these sequences in word-final stressed position, using the consonants [b, m, v, z, d, l, r] and vowels [a, u]. For each speaker and each C–V combination, only one repetition was analyzed.

C–V SEQUENCE TYPE	EXAMPLE WORD (in Russian)	EXAMPLE WORD (in transcription) 2	GLOSS
CV	тогда	tag'da	'then'
C^{j}V	судя	su'd^{j}a	'judging'
C^{j}jV	судья	su'd^{j}ja	'judge'
C^{j}ijV	судия	sud^{j}i'ja	'judge' (archaic)
CV	в углу	vu'glu	'in the corner'
C^{j}V	молю	ma'ɫ^{j}u	'I pray'
C^{j}jV	налью	na'ɫ^{j}ju	'I will pour'
C^{j}ijV	колею	kaɫ^{j}i'ju	'rut' (Acc. sg)

Table 1. Example words from the list of materials used in this study.

There are very few minimal pairs illustrating the contrast between palatalized-plus-jot and palatalized-i-jot. And, as in the contrast between [sud^{j}ja] and [sud^{j}ija], it is often the case that the palatalized-i-jot member of a minimal pair is archaic. The functional load of the other contrasts though is higher. For example, there are numerous pairs such as [sud^{j}a] 'judging' and [sud^{j}ja] 'judge', [s^{j}el] 'he sat' and [s^{j}jel] 'he ate', [l^{j}ot] 'ice' and [l^{j}jot] 'she/he/it pours'. (Bryzgunova, 1963)

Speakers read each word in randomized order and then repeated the final consonant–vowel sequence in isolation. We took measurements from these final isolated productions and also used a subset of them in the perception study. To justify our decision of using the isolated productions, we initially took measurements from both the words and isolated productions for one Russian speaker and found no difference in the results.

[2] Words are given in broad phonetic transcription.

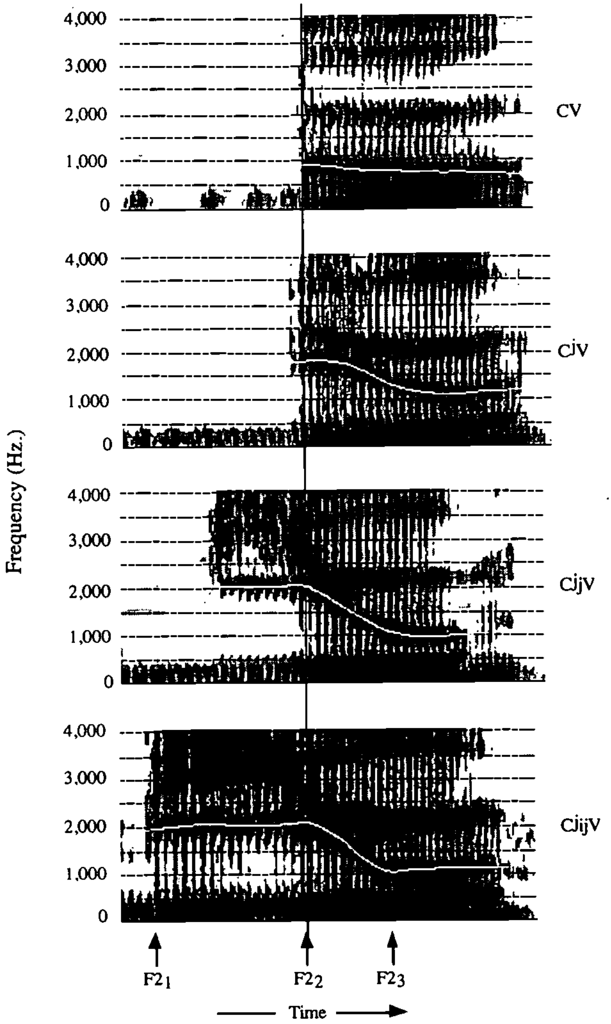


Figure 1. Sample spectrograms of the four C–V sequences. Spectrograms are aligned according to vowel onset as indicated by the continuous vertical line. The white lines were hand-drawn to show the center frequency. For each C–V sequence up to three F2 measurements (F2$_1$, F2$_2$ and F2$_3$) were made. The arrows below the last spectrogram (C^{j}ijV) indicate these three points of measurement.

We measured the frequency of the second formant at three points in time (see Figure 1). The first measurement ($F2_1$) was taken at the release of the primary occlusion of the consonant. Both the frequency of F2 and the temporal location of the release were noted. The second measurement ($F2_2$) was at the end of the relatively steady-state portion of F2, just prior to the transition to the following vowel. Finally we measured the F2 frequency and temporal location of the onset of the vowel steady-state ($F2_3$).

RESULTS

Figure 2 shows results for the male and female Russian and American speakers (see also Derkach *et al.*, 1970). The graphs show F2 trajectories for the palatalized, palatalized-plus-jot, and palatalized-i-jot sequences averaged across speakers, consonants and vowels.

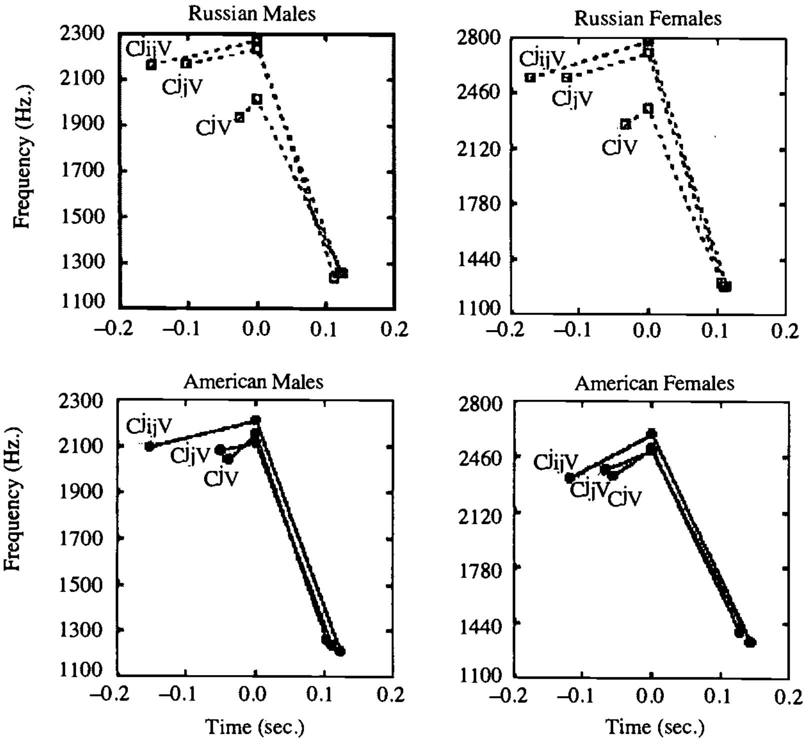


Figure 2. Results of the production study. Average F2 trajectories of the palatalized, palatalized-plus-jot, and palatalized-i-jot syllables are shown for each group of speakers. The trajectories are time aligned at the onset of the F2 transition to the following vowel.

For the Russian speakers, palatalized consonants had much shorter and lower frequency F2 steady-states than did either the palatalized-plus-jot, or the palatalized-i-jot sequences. Results of an ANOVA indicated that the steady-state durations are significantly different for both male and female Russian native speakers, [F(2,14)=173.1, p<0.01; F(2,14)=268.5, p<0.01, respectively] And, although the distinction between the palatalized-plus-jot and the palatalized-i-jot sequences is not large, a post-hoc Scheffe test indicated that there was a significant difference between the two, [for both male and female Russians all, p<0.01].

Turning now to the L2 learners, we see that the differences among the sequence types were much smaller than they were for the native speakers. Results of an ANOVA indicated that there was an overall significant difference in the F2 steady-state durations of the three C–V sequences of Americans' pronunciations for both males and females, [F(2,10)=81.1, p<0.01; F(2,10)=61.9, p<0.01, respectively]. However, an additional post-hoc Scheffe test indicated that there was not a significant difference in the F2 steady-state duration between the palatalized and the palatalized-plus-jot sequences, [p>0.05 for both males and females]. Results of this same post-hoc Scheffe test, however, did indicate that the Americans produced the palatalized-i-jot sequence with a longer F2 steady-state than in the other sequences, [p<0.01 for both male and female Americans].

In summary, this study showed that the L2 learners did not distinguish all of the palatalization contrasts that native speakers do, and that even though the distinction between palatalized-plus-jot and palatalized-i-jot does not carry much functional weight in Russian, native speakers do (at least in the speaking situation that we set up) maintain the distinction in production. Interestingly, the contrast that the L2 learners failed to produce is not the functionally weak contrast, but rather the more important (and perhaps less salient) contrast between palatalized consonants and the palatalized-plus-jot sequences.

PERCEPTION STUDY

METHODS

Forty-six listeners participated in the perception experiment. Eighteen (8 women, 10 men) were native speakers of Russian, and 28 (12 women, 16 men) were American learners of Russian. There was greater range of experience with Russian among the American listeners in this experiment, as compared with the Americans who participated in the production study. In a future report we will delve into the relationship between listeners' L2 experience and their performance in this study. In this report, however, we will discuss the American listeners as a group.

The speech samples presented to listeners were the productions of one (typical) male Russian speaker from the production study discussed above. We also included sequences with the vowel [i] in addition to [u] and [a]. The total number of stimuli presented to listeners was 252 consonant–vowel sequences, consisting of three repetitions of each of the 21 CV combinations of each of the four sequence types. The stimuli were presented in random order.

The listener's task was to identify each C–V sequence in a four-alternative forced-choice task. Prepared answer sheets listed each of the four sequence types (written in Russian) that were appropriate for a given C–V sequence. For example, if the test token was [b^{j}a] the alternatives listed were ба, бя, бья and бия ([ba], [b^{j}a], [b^{j}ja], and [b^{j}ija]).

RESULTS

Table 2 shows confusion matrices for both the Russian and American listeners. In these tables, stimuli and responses are coded according to sequence type. Data are summed over listeners, consonants and vowels. The stimuli are shown in the rows and listeners' responses are listed in the columns. For example, of the 1134 presentations of non-palatalized consonants the Russian native-speaking listeners labeled 1130 of them correctly as non-palatalized, and 4 times heard a non-palatalized token as 'palatalized'.

RUSSIANS

Stimuli	Responses			
	CV	C^jV	C^jjV	C^jijV
CV	**1130**	4		
C^jV		**1132**	2	
C^jjV		32	**1082**	16
C^jijV	1	10	826	**296**

AMERICANS

Stimuli	Responses			
	CV	C^jV	C^jjV	C^jijV
CV	**1575**	141	43	5
C^jV	152	**1288**	306	18
C^jjV	17	526	**908**	310
C^jijV	16	106	264	**1378**

Table 2. Results of the perception text. The confusion matrices show responses (in columns) to the four types of stimuli (rows) for native Russian speakers and American learners of Russian.

The cells that are shown in bold are the correct responses. If there were no responses for a given stimulus/response pair the cell is left blank. Note that overall the Americans showed greater confusion than did the Russian native-speakers. This is apparent in the fact that there are no blank cells in the American confusion matrix.

Because there were unequal numbers of Russian and American listeners it is convenient to present the confusions in percentages rather than raw counts (see Table 3). These data indicate that though the American speakers made no distinction between the palatalized and palatalized-plus-jot in production, they were able to distinguish them, though imperfectly, in the speech of a Russian native speaker.

RUSSIANS

Stimuli	Responses			
	CV	C^jV	C^jjV	C^jijV
CV	**99.6**	0.4		
C^jV		**99.8**	0.2	
C^jjV		2.8	**95.4**	1.4
C^jijV	0.1	0.9	73.0	**26.0**

AMERICANS

Stimuli	Responses			
	CV	C^jV	C^jjV	C^jijV
CV	**89.0**	8.0	2.4	0.3
C^jV	8.6	**73.0**	17.3	1.0
C^jjV	0.9	30.0	**51.5**	17.6
C^jijV	0.9	6.0	15.0	**78.0**

Table 3. Results of the perception test. The data shown in Table 2 are presented here as percentages rather than counts.

However, perhaps the most striking aspect of these data is that the Americans scored 78% correct for the palatalized-i-jot sequences while Russian native speakers correctly identified these sequences only 26% of the time. These data are unusual among studies of L2 acquisition because they show a contrast that is perceived more accurately by second language learners than it is by native-speakers.

GENERAL DISCUSSION

Previous research has found that near-mergers tend to be characterized by a pattern in which the contrast is lost in perception but is maintained in production (see Labov, 1994, p. 357). This is exactly the pattern that we found for Russian native-speakers in the contrast between palatalized-plus-jot and palatalized-i-jot in this study. Because the contrast is characterized by maintenance of distinctiveness in production, but loss of distinctiveness in perception, we conclude that this is an example of near-merger.

In the course of this study we have also observed that American speakers merge the Russian palatalized and palatalized-plus-jot sequences in production perhaps because they produced the palatalized consonants as a sequence of a non-palatalized consonant followed by a palatal glide. We have discussed this aspect of these data in more detail elsewhere (Diehm, 1996).

Finally, the Americans' performance in the perception experiment indicates a lack of awareness of the near-merger of the palatalized-plus-jot and the palatalized-i-jot sequences. They attended to a phonetic contrast without regard to its linguistic status. The Russian native-speakers, on the other hand, treated the difference (which the Americans' performance shows was perceivable) as disregardable variation in what is essentially one category encompassing both types of sequence.

These data suggest that L2 learners may attend at a psychoacoustic level to phonetic phenomena which are ignored by native speakers. Future research will tell whether we have uncovered a general characteristic of second language perception.

REFERENCES

Avanesov, R. I. (1972) *Fonetika sovremennogo russkogo literaturnogo jazyka*. Moscow: Iz. Moskovskogo universiteta.

Bolla, K. (1981) *A Conspectus of Russian Speech Sounds*. Hungary: Bölaj.

Bryzgunova, E. A. (1963) *Prakti eskaja fonetika i intonacija russkogo jazyka*. Moscow: Iz. Moskovskogo universiteta. pp. 50-53.

Derkach, M., G. Fant, and A. de Serpa-Leitao. (1970) "Phoneme Coarticulation in Russian Hard and Soft VCV-Utterances with Voiceless Fricatives." STL-QPSR 2–3. pp. 1-26.

Diehm, Erin. (1996) "Learning the Timing of Russian Palatalization." A paper presented at the meeting of the American Association of Teachers of Slavic and East European Languages (AATSEEL), Dec. 27-30, Washington DC.

Labov, W. (1994) *Principles of Linguistic Change: Internal Factors*. Oxford: Blackwell.

Panov, M. V. (1967) *Russkaja fonetika*. Moskva: Prosveščenie.

Logan, J. (1988) "The identification of speech using word and phoneme labels." Research on Speech Perception, PR13, Speech Research Laboratory, Department of Psychology, Indiana University. pp. 277-306.

Zinder, L. R., L. V. Bondarko, and L. A. Verbitskaja. (1964) "Akustičeskaja xarakteristika različija tverdyx i mjagkix soglasnyx v russkom jazyke." Učenye zapiski LGU. Serija filologičeskix nauk, vyp. 69. pp. 28-36.

OSU Working Papers in Linguistics 50 (19-39)

Breadth of focus, modality and prominence perception in Neapolitan Italian[*]

Mariapaola D'Imperio
dimperio@ling.ohio-state.edu

Abstract: This study explores the notions of "nuclear stress", "accent placement" and "breadth of focus" in the Neapolitan variety of Italian. The predictions of standard generative theories about their interrelationships are tested through a perceptual study employing statements and questions with varying focus structure. The results show that broad focus statements are more ambiguous than late narrow focus ones as to the extraction of intended focus pattern. Broad focus questions are, in turn, less ambiguous than broad focus statements for the same purpose. The results suggest the importance of the role of accent type differences.

INTRODUCTION

The goal of the present study was to test two claims about the relationships among prosodic prominence, intonational accent and scope of focus that are standardly assumed in generative linguistic theories. The language investigated is the variety of standard Italian spoken in Naples. Two experiments with forced choice response were designed to assess different aspects of this complex of phenomena. Experiment I was intended to elicit responses that indirectly measured location of prominent words, while Experiment II explored focus scope more directly through a question-answer matching paradigm. For both experiments, speech stimuli varying in focus placement (early, medial, late) and focus type (broad vs. narrow) were employed.

Theoretical claims about the relationship between accent and stress on one side, and between focus and accent/stress on the other will be reviewed in order to formulate hypotheses and predictions. Therefore the next section will be concerned with these claims and with possible implications for the present study.

SENTENCE STRESS, NUCLEAR ACCENT AND FOCUS OF INFORMATION

In describing English intonation, two main themes recur in the literature, namely the notion of "sentence stress" and the notion of "focus" and its scope. I shall first try to clarify the former notion as it has been interpreted in various theoretical frameworks.

Sentence stress was defined in the earliest versions of Generative Phonology as the derivational product of the "nuclear stress rule" as formulated by Chomsky and Halle (1968). In neutral utterances, the application of this rule renders the last stress of a phrase the most prominent, that is the "nuclear" stress of the utterance. The nuclear stress corresponds to the "tonic syllable" in works such as Halliday (1967) and to the "nuclear syllable" in Crystal (1969). In these theories, sentence stress was recognized to be inextricably tied to the intonation pattern; the nuclear stress was defined in terms of the

[*] This work represents the partial results of a research that I started last summer, thanks to the Cognitive Science Department Summer Fellowship, which was inspired by my conversations with Prof. Mari Jones. Special thanks to Prof. Mary Beckman for extensive guidance in the whole process; thanks also to Mike Calcagno, Craige Roberts and all the members of the "Pragmaticats" group for helpful suggestions regarding Experiment II.

location of the only prominent pitch event or "accent" of the intonational phrase. Pierrehumbert (1980) and subsequent works by her colleagues changed this view, in that in her theory the nuclear accent is not necessarily the unique prominent melodic event of the intonational phrase; rather, it is one of the potential accents of the phrase (though it is characterized by having a special status).

In the original Pierrehumbert's theory, which was cast in a derivational framework, stress assignment precedes accent assignment in that "pitch accents are lined up with the text on the basis of the prominence relations" (Pierrehumbert 1980, p.102). Therefore, a pitch accent must occur on a syllable that has already been designated by the grammar as being metrically the strongest, or most "stressed", in that phrase. Selkirk (1984) embodies another version of the traditional derivational approach, but in her framework accent assignment has to precede stress assignment. More recent approaches suggest that we do not need a serial process of stress-to-accent or accent-to-stress application, but we can think of various constraints acting in parallel to produce a well-formed output (Pierrehumbert 1993; Beckman, 1996).

The original theory was couched in terms of Liberman and Prince's (1977) account of stress as metrical (i.e. rhythmic) strength. It stated that the nuclear accent was associated to the D.T.E. (Designated Terminal Element) syllable, and that earlier accents could be associated only to relatively strong syllables, at grid levels no higher than the nuclear accented syllable. Beckman (1986) proposed a different account of metrical strength, which incorporated the results of phonetic experiments such as Fry (1958). It is this more intonationally "direct" account of stress that is incorporated in the ToBI conventions (Beckman & Ayers, 1994). In this account, stress is envisaged as a hierarchy of prominence levels, where each level is defined in terms of its own phonetic properties. It is useful to think of this idea in terms of the following grid model:

```
4.                  +       nuclear accent
3.+                 +       pitch accent
2.+       +         +       full vowel
1.+   +   +     +   +       syllable
 Ronnie loves Marie
```

Fig. 1 Grid representation of the utterance *Ronnie loves Marie.*

According to this view, there are three degrees of stress prominence above the syllable level, as represented in figure 1 by level 2, 3 and 4. First, this representation allows us to talk about a hierarchy of qualitative distinctions between accented vs. unaccented (though stressed) syllables. For instance, in figure 3 the syllable *loves* is marked by the first level of stress (level 2) because of the nature of its vowel, which is "full" and not reduced, unlike the vowel of the first syllable in *Marie*; however, this syllable is not marked by the second level of stress (level 3), i.e. it is not (pitch-)accented. We must underline that according to this theory of intonational phonology, an accented syllable is the product of the association of a pitch accent with a metrically strong syllable. The third and final level of stress (level 4) defines the nuclear accented syllable, which is the most stressed among all possible accents in the phrase. As we can notice from the grid representation in figure 1, the strongest stress (level 4) is located where prominence, marked by the presence of a cross at each level, reaches the highest point. The pattern that traditional generative theories might have introspected as having "normal stress" is the one depicted above. In these circumstances, the nuclear accent is the last, and considered to be the most prominent pitch accent in a phrase, whereas a pre-nuclear accent is any pitch accent preceding the nuclear one (see the grid mark on *Ronnie* at level 3 in Fig. 1). This representation entails a notion of stress as a complex, structural entity, while previous notions of stress, typical of older accounts, view stress as a separate, "suprasegmental" feature (see Lehiste, 1970).

It is also clear that prominences can be controlled in order to obtain specific focus effects. A word or a phrase can be singled out by the nuclear accent placement as the most important bit of information that the speaker wants to convey. Accent placement is in fact clearly related to focus of information. The notion of nuclear stress played an important role in the interpretation of the notion of "focus" in studies of generative syntax and semantics of the late 1960s and the 1970s (Chomsky, 1971; Jackendoff, 1972; Selkirk, 1984). In this framework, the nuclear stress was invoked in order to promote the last stress of a "neutral" utterance to the nuclear stress. A neutral utterance is the default case, in which no particular word or phrase in the utterance is singled out as being the most important or the "new information" to convey. The nuclear stress rule was then said to produce "normal stress".

After a long debate about this notion of "normal stress", Ladd (1980) reinstated the validity of the pattern and coined a term for it, that is "broad focus", which is still used today to refer to "focus on whole constituents or whole sentences, not just on individual words" (Ladd 1980, p.3). In this perspective, normal stress describes accent placement when focus is broad. "Narrow" focus is the term that naturally opposes "broad" focus and refers to cases where smaller constituents such as single words are selected as the focused element. One way to think of the notion of focus scope is by considering felicity of a statement as an answer to a particular question. For example, statement B in (1), where the constituent under focus is made of a single word, might be uttered in the context of question A (the word in capital letters indicates the nuclear accented word within the focused constituent marked with brackets).

(1) A: "Who loves Marie?"
B: "[RONNIE] loves Marie"

(2) A: "What does Ronnie feel?"
B: "Ronnie [loves MARIE]"

However, in (2), where the context is represented again by question A and the answer is B, the constituent under focus is larger, i.e. it is the whole verb phrase.

One of the additional claims of the standard theory of the relationship between focus and accent is that the phonological form of an utterance with late accent placement will be ambiguous between a broad focus interpretation and a late narrow focus interpretation (Chomsky, 1971; Jackendoff, 1972). For instance, a pattern like "Ronnie loves MARIE" is potentially ambiguous between late narrow focus and broad focus readings. In other words, this utterance can be the phonological expression of all of the focus structures in (3).

(3) a. "[Ronnie loves Marie]$_F$"
b. "Ronnie [loves Marie]$_F$"
c. "Ronnie loves [Marie]$_F$"

In (3a) the whole sentence is in focus, in (3b) focus is on the entire verb phrase, and (3c) has narrow focus on the noun.

Accent placement works very similarly in Italian. As in English, the location of a pitch accent associated to a stressed syllable is not fixed in a prosodic phrase. For instance, in a Subject-Verb-Object declarative sentence, accent can be placed on any of the words, according to which of them is in focus. In the sentence *Giovanni ama Maria* "John loves Maria", all of the prominence patterns in (4) are possible.

(4) a. *GIOVANNI ama Maria*
b. *Giovanni AMA Maria*
c. *Giovanni ama MARIA*

While in (4a) the nuclear accent is early, in (4b) it is medial and in (4c) it is late. According to the standard theory of the relationship between focus and phrasal phonology mentioned above, the accentual pattern of (4c) is ambiguous, since it could signal narrow focus on the object, broad focus on the verb phrase or even broader focus on the sentence as a whole (i.e. completely "new" information). This ambiguity represents a problem for a theory that would directly derive focus from accent placement, and is caused by the lack of a one-to-one correspondence between focus and accent. The question is, then, from the view of perception: "How do listeners determine the breadth of focus, given a certain accent placement?". Under a theory that privileges the role of syntax, such as Selkirk (1984), one derives focus interpretation via argument structure and "percolation" rules that allow the [+F] ("focus") feature specification to percolate up the syntactic tree. Other theories would privilege the semantic contribution and the difference in "informativeness" of broad focused and narrow focused words (see Ladd, 1996, chap. 5, for a review of some other accounts that have been proposed to accomodate them).

The present study tests the predictions of standard Generative Phonology, namely whether late and broad focus are actually perceptually confusable. If the claims of standard phonological theories are true, we predict that utterances with late accent placement, whether having an intended broad focus structúre or having late narrow focus, will be equally perceptually ambiguous.

ACCENT STATUS AND PROMINENCE

The typical prominence structure of a broad focus utterance can be thought of as having the same representation of the utterance in figure 1 (above). When Ladd (1980) formulated the notion of broad focus, though, there was still no clear notion of prenuclear (as opposed to nuclear) accent, since the only accent of the intonation phrase was identified with the nuclear one. However, we know that one of the constraints of the phonology on the prosodic shape of the utterance is that content words preceding the nuclear accented word can often be accented too (see prenuclear accent in figure 1). As mentioned above, current theories of intonational phonology assume that when more than one accent is present in a phrase, it is the last one that will be associated with the designated "most stressed" syllable in the phrase.

Under such a hierarchical view, the prediction is that nuclear accented words will be more prominent than pre-nuclear accented words, which in turn will be more prominent than unaccented words. In this perspective, different types of nuclear accent will be characterized by the same degree of prominence or at least by degrees of prominence that do not reverse the pattern of relative prominence of the prenuclear and nuclear stress within the same phrase. For instance, the nuclear accent typical of English statements is H*, while a "downstepped" H* (transcribed as !H*; see Beckman and Ayers, 1994) tends to be used in a more narrative type of tune and is characterized by a lower fundamental frequency (F0) peak. However, both nuclear accent types will produce equal degrees of perceptual prominence, in that they are attributed an equal status in the stress hierarchy. Previous works on prominence perception had suggested that listeners are able to normalize for the natural F0 declination during the course of an utterance (Pierrehumbert 1979), in that they would perceive a later H* as being as prominent as a preceding H*, even though (all other things being equal) the late H* has necessarily a lower F0 peak than the preceding H*. Work by Horne (1991) and Ayers (1996), on the other hand, suggests that a nuclear syllable bearing a downstepped !H* accent may not be as fully prominent as a syllable with a H* in the same position. The observation is that accent type differences can contribute to the perception of more or less strong prominences.

Accent type differences can be observed also in Italian. The nuclear accent type commonly found in statements with broad focus is characterized by a rather downstepped quality. Figure 3 shows two different productions of the sentence *Mamma andava a ballare da Lalla* "Mom used to go dancing at Lalla's". The upper panel presents an example of broad focus, with a tune transcribed as H* H+L* L-L%. What is important to notice here is the relatively shallow F0 variation within the nuclear accented syllable marked with H+L*, as opposed to the greater F0 excursion within the prenuclear accented H*. Narrow focus statements differ from broad focus ones in that they present a nuclear pitch accent that is acoustically more salient, and are characterized by a H*+L nuclear accent (lower panel).

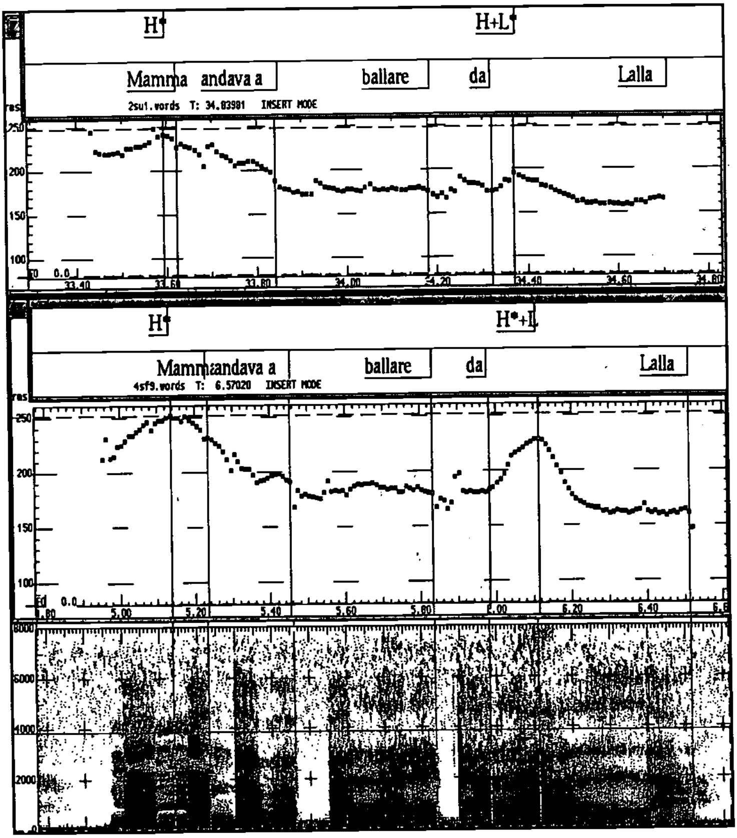


Fig. 2 Broad focus (upper) and late narrow focus (lower) realization of the sentence *Mamma andava a ballare da Lalla* uttered as a statement.

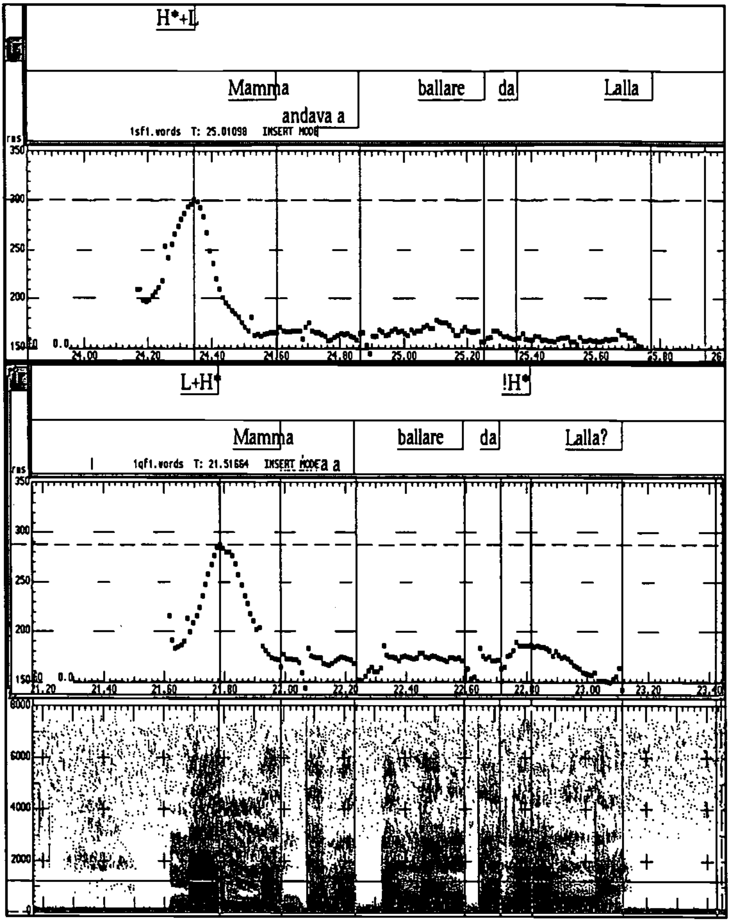


Fig. 3 Pitch contour of *MAMMA andava a ballare da Lalla* "Mom used to go dancing at Lalla's" uttered with early nuclear accent on *mamma* both as a statement (upper panel) and as a yes/no question (lower panel).

Another important defining feature of the nuclear accent in Pierrehumbert's framework is that it is the pitch accent that immediately precedes the phrase accent in English. A phrase accent is the tonal event that controls the pitch level in the region that goes from the end of nuclear accented word up to the end of the phrase. The role of the phrase tone cannot be underestimated, since it is through this event that we can relate the nuclear accent of English to that of languages like Swedish. While in English the phrase accent prevents the realization of post-nuclear accents, in Swedish pitch accents occur beyond the phrase accent, but are downstepped (Bruce 1982).

The suggestion that downstepped accents might mark less prominent syllables is especially intriguing when we compare the role of downstep in Swedish and in Neapolitan Italian. While early focus declaratives in both the standard and the Neapolitan variety of Italian present the same characteristics of English early narrow focus utterances (by showing the predicted flat melodic configuration following the nuclear accent), yes/no questions of the Neapolitan variety do not. In this variety, yes/no questions with early focus (see Fig. 3 above) present a sharp pitch obtrusion to a pronounced peak (L+H*) on the focused constituent, plus a smaller peak on the last stressed syllable of the intonational phrase (!H*).

To sum up, the standard accounts of nuclear stress in English can be understood as a purely "positional" definition. The nuclear stress is the syllable in D.T.E. position. It is the syllable positioned to hear the last pitch accent in the intonation contour (in English), or the accent positioned just before the phrase accent (in English or Swedish). The common thread underlying all of these definitions is that sentence stress (as represented, e.g., in the grid) is independent of accent type, even if it is not independent of intonation, as assumed in early Generative Phonology. I will refer to this hypothesis as the "positional hypothesis". Moreover, the traditional generative view of the relations between the pragmatics of focus and accent placement claims that an accent placed on the last element of an utterance can ambiguously signal a broad or a late narrow scope of the focus. By the same token, the accent structure of stimuli with either (intended) broad or late narrow focus will be equally perceptually confusable. I will refer to this as the "(late accent) ambiguity hypothesis".

The results of the experiments presented in this paper appear to provide evidence against the ambiguity hypotheses, but we could not find conclusive evidence for the stress-to-prominence relation claimed by the positional hypothesis. Moreover, none of the hypotheses considered could predict the unexpected difference in the patterning of results between questions and statements found here.

EXPERIMENT I

METHODS

Stimuli

The stimuli consisted of a set of sentences with different word number, focus pattern and modal intonation (question vs. statement). Four groups of sentences, with four sentences in each group, were created. Group I consisted of three-word sentences, uttered as statements; group II consisted of two-word sentences, uttered as statements; groups III and IV were uttered as questions, and included respectively three-word and two-word sentences. All of the sentences had either an SVO (Subject-Verb-Object, e.g. *Maria ama Giovanni* "Maria loves Giovanni") or SV (Subject-Verb, e.g. *Mario esce* "Mario goes out") structure, depending on the number of words in the sentence. The words employed had a variable number of syllables and of lexical stress pattern (initial vs. non-initial).

As shown in Figure 4 (below), each of the sentences for each type was uttered as either a neutral utterance with broad focus (focus B) or as a narrow focused utterance, where scope of focus was limited to a single word - either S (focus S), V (focus V) or O (focus O). The set of sentences was produced by two speakers of Neapolitan Italian (the author and a male speaker), each of whom read half of it. The production of the male speaker was strictly monitored by the author, and the sentences were all auditorily transcribed to check for intended focus pattern.

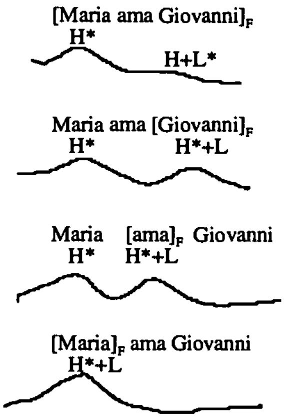


Fig. 4 Stylized statement contours with different focus structure.

The recordings were made in the Department of Linguistics Lab, Ohio State University, where they were digitized at 16 kHz on a SUN Sparc Station using ESPS Waves+. The stimuli were coded and pseudo-randomized using a Latin square design. In this arrangement each treatment occurs just once in each row and just once in each column. It was ensured that each of the speakers (the female and the male voice) in the test would not produce two consecutive treatments in each row. The stimuli were then recorded on a tape, which was later played to the subjects.

Procedure

The test was administered to a homogeneous group of participants, consisting of 20 university students and recent graduates, with ages varying from 23 to 29. From the original 22 participants, the data relative to 2 of them had to be discarded because their response pattern suggested that they were not attending to the task. The participants were all speakers of Neapolitan Italian and hence had the same linguistic background as the speakers that produced the stimuli. Only three of the participants were knowledgeable in Linguistics, but none was aware of the purpose of the experiment.

The method used to gather perceptual data was as follows: each participant was given a reply sheet on which 56 sentences were written, each consisting of either two or three words, separated from each other by means of framed boxes.

Ex.

Mario	esce

Maria	ama	Giovanni

The set of sentences was divided into four blocks. The participants listened to each sentence and had to mark only one of the words in the sentence, with a check or a cross on a specific box; specifically, they were asked to mark the word that appeared to be the most "important" in the sentence. Subjects were also told to mark the answer as quickly as possible after listening to each stimulus, even when not entirely sure about the answer. Use of linguistic terms such as "prominence" and "focus" was avoided and the attribute used throughout the experiment was "important". It was our concern to leave linguistic notions outside of the task, so that even naive listeners could perform it without confusion.

A short training session preceded the set of trials, where the experimenter presented examples of utterances with varying intended focus structure (see Figure 4 above) and had the participant point at one of the words as being the most important. Each sentence was separated from the following by a three-second pause. To facilitate the task, a short beep separated each block of utterances from the other. The beep was inserted to indicate the stage of the experiment reached, while knowing that the fourth beep would signal the end of the experiment itself.

Design and predictions

Two factors were included in the experimental design. The focus factor, derived from labeling the intended focus pattern following the standard theory, and the modality factor (question vs. statement intonation), will be referred to respectively as FOCUS and MODALITY. FOCUS has 3 levels in the two-word sentences (focus B, focus S and focus V) and 4 levels in the three-word sentences (focus B, focus S, focus V and focus O); MODALITY has always 2 levels (question modality vs. statement modality). Therefore, the design was a 2 (modality) x 3 (focus) factorial in two-word sentences and a 2 (modality) x 4 (focus) factorial in three-word sentences. Both independent variables, FOCUS and MODALITY, were manipulated within subjects.

For two-word stimulus sentences, responses were transformed into the dependent variable PERCENT, which is the percentage of responses which assigned "importance" to the Verb for each stimulus token. We will assume for now that "important" is equal to "prosodically prominent", and on the basis of this assumption we will make the following predictions. According to the positional hypothesis, we expect that in focus B and focus V stimuli, where the word carrying the nuclear accent is the Verb, we will have a mean percent of "assigned importance" of around 100%, whereas the value of percent "importance" assigned to the Subject will be close to 0%. By the same hypothesis, we predict that the value of assigned importance to the Verb will be around 0% in focus S stimuli, while assigned importance to the Subject will be close to 100%. By the ambiguity hypothesis we also predict that, whichever the exact value will be, focus B stimuli will have an exact same value of assigned importance to the Verb as focus V stimuli, since they both have late accent placement.

For three-word stimuli, we have to set three independent variables, since participants had a choice among three and not just two elements for the purpose of assigning importance. The three variables were obtained by calculating percentages of assigned importance to each of the three word positions, i.e. Subject, Verb and Object, yielding respectively the variables PERCENT-S, PERCENT-V and PERCENT-O. Again, according to the positional hypothesis, we expect that, independent of breadth of focus, in all stimuli types (focus B, focus S, focus V and focus O), the word carrying the nuclear accent will have a mean percent of "assigned importance" of around 100%, while the other words will receive values of assigned importance close to 0%. This means, for instance, that the value of PERCENT-S for focus S stimuli will be around ceiling. By the ambiguity hypothesis we again predict that, whichever the exact value, focus B stimuli will have the same value of assigned importance to the Object as focus O stimuli, since they both have late accent placement.

Two-way ANOVAs were performed on the dependent variables. Post-hoc tests (Tukey Kramer) were performed to test the relative significance of the mean differences in all cases of significant interaction. The criterion was set to p<.05 for both planned comparisons and post-hoc tests.

RESULTS

Two-word utterances

As expected from the predictions of the positional hypothesis, the highest value of PERCENT-S (95.63%) was found in utterances with narrow focus on the subject. An interesting result is represented by the pattern shown by broad focus utterances, where, contra the predictions, we find a value of 35.63% for PERCENT-S. What is more, utterances with intended focus on the verb received assigned importance on S in 14.38% of the cases, against expected values of around 0%.

In figure 5 the interaction data for both questions and statements are presented. Here, PERCENT-S is plotted on the y axis across the three utterance types (focus S, V and B). Contrary to the expectations, statements with focus on the verb have a mean of 22.5% responses which assigned importance on the subject. Two results were not predicted by either of the theories mentioned above. First, questions received lower overall values for PERCENT-S than statements. Second, there was a conspicuous difference between broad focus questions and statements. While the category of broad focus utterances produced, overall, a high PERCENT-S value, the biggest contribution is due to broad focus statements, which averaged 55% PERCENT-S (as compared to 36.5% PERCENT-S for questions).

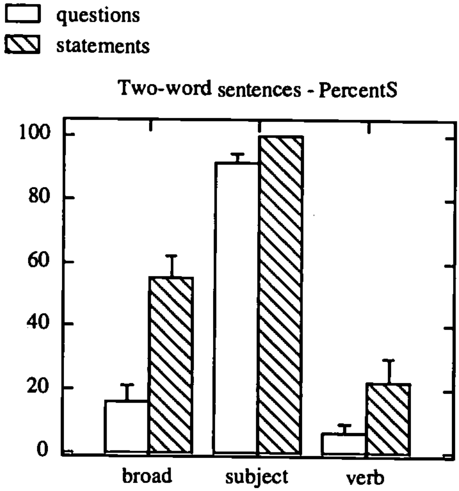


Fig. 5 PERCENT-S values by FOCUS (Broad, Subject, Verb) and MODALITY (question vs. statement) for two-word utterances.

Both FOCUS [F (2,18) = 136.0; p<.05] and MODALITY [F (1, 18) = 25.924; p<.05] were found to significantly influence the pattern of responses. A significant interaction between the independent variables was also found [F (2, 18) = 4.664, p<.05]. A post-hoc comparison revealed a significant difference between all the levels of the first and the second independent variable. Therefore, the observed trend of utterance modality (question vs. statement) in determining the percentage of assigned importance was proved to be significant.

Three-word utterances

Narrow focus utterances

Figure 6 shows the pattern of PERCENT-S responses for three-word sentences. In utterances with narrow focus, most subjects judged the focused word as most important, without regard to modality, with 94% of responses judging the Subject as "most important" in focus S utterances, 87% judging the Verb as most important in focus V utterances and 77% judging the Object as most important in focus O utterances. Again, with the assumption that prominent equals "important", we can make various observations. In a way that was not predicted by the positional hypothesis, utterances with narrow focus on the object received a fair amount of assigned importance on the subject, i.e. 17%, and a minor percent of assigned importance on the verb, i.e. 6%. In utterances with narrow focus on the verb, just 6% of the subjects judged the intended focus to be on the subject and 7% to be on the object.

Figure 6 reveals also that modality, i.e. question vs. statement intonation, affects the pattern of utterances with narrow focus on the verb. In particular, questions produce a percentage of 79% for this intended focus structure, whereas statements yield a value of 95%. This trend can be found, to a lesser extent, in utterances with focus on the subject; within this category, questions produced 89% of PERCENT-S responses, against 99 % for statements.

As for the statistics on the values of PERCENT-S, both FOCUS [F (3, 24) = 71.220; p<.05] and MODALITY [F (1, 24) = 4.854; p<.05] resulted significant. Similarly, an interaction of the independent variables was found [F (3,24) = 3.523; p<.05]. Post-hoc tests revealed significant differences between utterances with focus S and every other focus level. The ANOVA performed on data for PERCENT-V revealed a significant effect of FOCUS [F (3,24) = 136.00; p<.05], once again, but neither a significant difference of MODALITY [F (1,24) = 1.05; p>.05] nor a significant interaction [F (3,24) = 2.235; p>.05]. A post-hoc comparison uncovered significant differences between utterances with focus on the verb and every other focus level. The data of assigned importance to the object will be discussed below.

Broad focus utterances

For broad focused utterances, the predictions of the positional hypothesis are not met. In fact, we find that only 56% of the responses assigned importance to the object within this focus category, as opposed to the very high value of this percentage for utterances with narrow focus on the object. Once we assume that importance equals prominence, the prediction of current phonological theory is that these two percentages should be roughly equal, since there should be no phonological prominence difference between utterances with late accent placement, whether they be narrowly or broadly focused. It was also found that the subject and the verb received a fair amount of assigned importance when focus was broad, in the measure of 26% and 18% respectively.

Even more interestingly, when looking at figure 6, we notice again that broad focus utterances pattern differently according to modality. First, the interaction pattern is reversed relative to what we saw previously. Overall, statements present smaller percentage values of assigned prominence to the object than questions. Remarkably, broad focus statements present just 35% of assigned importance to the object. As observed above, statements with narrow focus on the object presented instead a much higher value, i.e. 77.5%. Questions, on the contrary, produced an equal percentage of responses for both of these focus levels, (76%), giving support to the ambiguity hypothesis. Moreover, broad focus statements present a very high and unexpected percent of assigned importance to the subject (43.74%), whereas broad focus questions produced a much smaller value (8.75%). The data relative to broad focus statements also seem to possess the greatest variability when compared to data for other focus structure types.

The analysis of variance on the percent of assigned importance to the object uncovered a significant effect of FOCUS [F (3,24) = 47.977; p<.05] and an effect of MODALITY [F (3,24) = 6.691; p<.05]. The interaction was also significant [F (3,24) = 3.207; p<.05]. A post-hoc test on the focus effect revealed a significant difference between focus B and focus S, focus B and focus V, focus O and focus S and focus O and focus V utterances. A post-hoc analysis performed on the modality effect resulted in a significant difference between questions and statements.

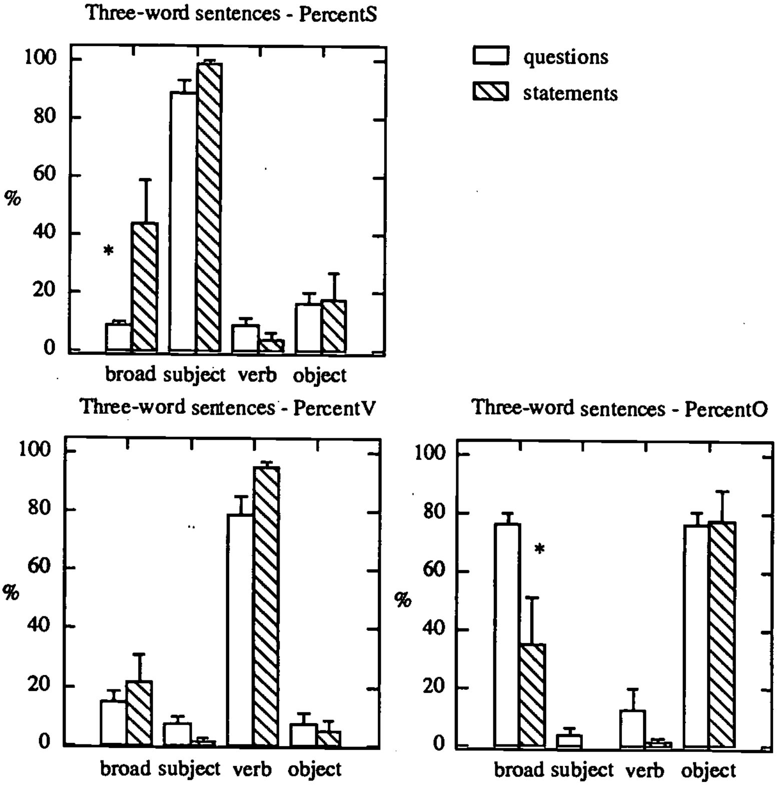


Fig. 6 PERCENT-S (upper), PERCENT-V (medial, left) and PERCENT-O (medial, right) values by FOCUS (Broad, Subject, Verb) and MODALITY (question vs. statement) for three-word utterances.

Therefore, contrary to the predictions of current phonological theory, a significant difference was found in the patterning of the data between utterances with broad focus and narrow focus on the object. As a relevant and novel result, while focus B and focus O questions produced identical values of PERCENT-O, focus B and focus O statements did not.

DISCUSSION

Experiment I attempted to assess the percept of prosodic prominence through the notion of "importance". However, the results are difficult to interpret in this light, and seem to suggest that listeners were responding in terms of focus structure, and not simply in terms of accent structure. Moreover, while a traditional view of the relationship between accent placement and focus structure (the ambiguity hypothesis) predicted that late accent utterances in general would show identical responses for both broad and late narrow focus, our results show that this is true only for questions.

The most interesting result of Experiment I, in fact, comes from the analysis of broad focused utterances as a function of modal intonation. Specifically, while statements with intended broad focus (focus B) obtained values of assigned importance that are significantly different from those obtained for statements with intended late narrow focus (focus V for two-word stimuli and focus O for three-word stimuli), the converse was true for questions.

To sum up, the results of Experiment I suggest that the listeners relied on the focus structure of the utterance more than on its prominence structure (given by the relation between nuclear and prenuclear accent in a prosodic hierarchy of accent status) in order to perform the "assigned importance" task. This unexpected outcome urged us to perform a second experiment where the notion of focus will be investigated more directly, in order to see to what extent listeners were responding in terms of focus in Experiment I and not in terms of accent relationships. However, Experiment II will still leave unanswered the question relative to prominence perception, which will probably need the use of a different experimental methodology. An alternative hypothesis for the unexpected results of the broad focus statements is that listeners were simply confused when it came to assigning "importance" to a word in an utterance with no prior context. Experiment II will also try to address this concern.

EXPERIMENT II

METHODS

Stimuli

The same stimuli as for Experiment I were employed (see the Stimuli section for Experiment I).

Procedure

The test was administered to a homogeneous group of participants, consisting of 23 university students and recent graduates, with ages varying between 22 and 29, and who were all speakers of the Neapolitan variety of Italian. Participants had to match the statement they heard to one of several questions setting up a specific context:

Ex. stimulus: *Maria ama GIOVANNI* "Mary loves JOHN"
matching question: *Chi ama Maria?* "Who does Maria love?" (focus O)
other choices: *Chi ama Giovanni?* "Who loves John?" (focus S); *Che pensa Giovanni di Maria?* "What does Maria think of John?" (focus V) or *Dimmi qualcosa di quella coppia* "Tell me something about that couple" (focus B).

When the stimulus heard was a question, the set of context-setting choices was very similar to the one used for statements, but in this case the participants were instructed to choose the question whose meaning most resembled that of the stimulus question, or that was felicitous as an utterance following what they just heard. For instance, in a dialogue, the question *Chi ama Maria* "Who loves Maria? " is felicitous in the context of an immediately following yes/no question *GIOVANNI ama Maria?* "John loves Maria? ".

Participants first listened to the test sentence and then chose the context question.

Design and predictions

The experimental design was the same as for Experiment I. The only difference was in the number of dependent variables on which the statistics was conducted. In Experiment II, the participants had to decide between three possible matching questions for two-word utterances and four for three-word utterances. The coding devised for each question typology was the following: PERCENT-B stands for percent of responses choosing the question congruent with a broad focus utterance, while PERCENT-S, PERCENT-V and PERCENT-O refer respectively to percent of responses choosing questions congruent with a narrow focus utterance (either focus S, focus V or focus O).

As in Experiment I, given the hypothesis that participants were responding in terms of focus structure, we expect very high values of PERCENT-S, PERCENT-V and PERCENT-O responses for utterances with narrow focus on the respective element. Especially high values are expected for questions and two-word utterances. Since the participants had to match a specific stimulus to a matching context-question, we also expect them to be more consistent in the pattern of responses than in Experiment I. This factor could especially be relevant for broad focus utterances, which produced highly variable responses in Experiment I. Our hypothesis can be formulated as follows:

Hypothesis 1: following the ambiguity hypothesis of traditional focus-to-accent accounts, we expect ambiguity of focus extraction for focus B and late narrow focus utterances (focus V for two-word stimuli and focus O for three-word stimuli). We must recall that this ambiguity was suggested by the results of Experiment I to be very high in question stimuli. Therefore, we expect high values of PERCENT-V for two-word broad focus utterances, where late accent placement will simply be confused with late narrow focus, and, by the same token, high PERCENT-O values for three-word broad focus utterances. This hypothesis predicts also that high values of PERCENT-S and PERCENT-V will be found for early narrow focus utterances, while late narrow focus utterances will receive equal percentages of PERCENT-O and PERCENT-B. This hypothesis is based on the claim that accent placement is what renders broad focus ambiguous and virtually indistinguishable from late narrow focus.

On the other hand, given the results of Experiment I, if hypothesis 1 is not completely supported it will be that broad focus statements will not necessarily be identified with late narrow focus statements, while broad focus questions will. Broad focus statements will instead present high values of PERCENT-B responses. Given the value of assigned importance to the Subject for broad focus statements in Experiment I, it could also be the case that broad focus statements will receive focus S responses in a few cases. If this scenario is confirmed by the results of Experiment II, we will have to find a suitable explanation.

RESULTS

Two-word utterances

Narrow focus utterances

The results clearly replicated those of Experiment I. The highest value of PERCENT-S (93%) was found in utterances with narrow focus on the subject, while the highest value of PERCENT-V (81%) was found in utterances with narrow focus on the Verb.

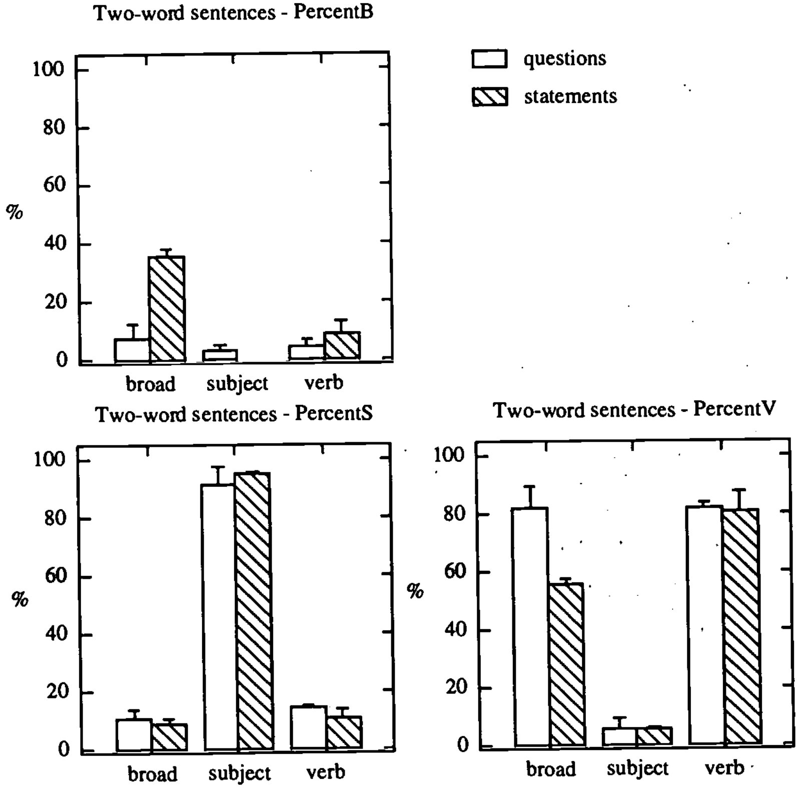


Fig. 7 PERCENT-B (upper), PERCENT-S (medial, left) and PERCENT-V (medial, right) values by FOCUS (Broad, Subject, Verb) and MODALITY (question vs. statement) for two-word utterances.

The ANOVA performed on PERCENT-S yielded the following results: FOCUS affected significantly the pattern of responses [F (2,18) = 453; p<.05], while MODALITY [F (1,18) = .80; p>.05] and the interaction between the variables [F (2,18) = .620; p>.05] did not. As for PERCENT-V, FOCUS was again highly significant [F (2,18) = 147; p<.05], but there was also a significant effect of MODALITY [F (1,18) = 5.531; p<.05] and a significant interaction [F (2,18) = 4.894; p<.05].

In figure 7 the interaction data for both questions and statements are presented. Here, PERCENT-B, PERCENT-S and PERCENT-V are plotted on the y axis across the three utterance types (focus S, V and B). Contrary to the expectations, statements with focus on the verb have a mean of 10% responses which assigned importance on the subject. However, this value is lower than the one reported for the same focus type in Experiment I. Also, differently from Experiment I, questions did not receive a lower overall value of PERCENT-S responses.

Broad focus utterances

Figure 7 shows also results for broad focus utterances. The results of Experiment I are replicated also here. The value of PERCENT-V is once again lower than the value reported for utterances with narrow focus on the Verb. As in Experiment I, broad focus statements receive a lower score of PERCENT-V than late narrow focus ones, even though the value is well above chance here (55%), which appears to support hypothesis 1. Also, the value of PERCENT-S is much smaller than the one reported for Experiment I (around 10%). Broad focus and focus V questions receive same values of PERCENT-V, strongly supporting hypothesis 1.

As to the values of PERCENT-B, they are highest for broad focus utterances, as expected, even though this value is relatively low (22%, below chance). Figure 7 reveals also a highly significant interaction [F (2,18) = 13.471; p<.05] between the two modalities. The value of PERCENT-B is due mainly to the effect of broad focus statements, which alone present a PERCENT-B value that is slightly above chance (36%). FOCUS was also significant in this case [F (2,18) = 21.902; p<.05], as well as MODALITY [F (2,18) = 14.294; p<.05]. A post-hoc comparison revealed a significant difference between questions and statements, as well as between broad focus and the other two focus types. The results confirm the tendency for broad focus questions to be identified with late narrow focus which was already noticed in Experiment I.

Three-word utterances

Narrow focus utterances

Figure 8 shows the pattern of PERCENT-S, PERCENT-V and PERCENT-O for three-word sentences with narrow focus, which replicated the results of Experiment I. In utterances with narrow focus, most subjects judged the focused word as most important, without regard to modality, with 92% of responses judging the Subject as "most important" in focus S utterances, 76% judging the Verb as most important in focus V utterances and 73% judging the Object as most important in focus O utterances. Differently from Experiment I, utterances with narrow focus on O were not incorrectly identified with focus S utterances in many cases (only 7%), while they were identified with focus V utterances with a slightly higher percentage (17%). In utterances with narrow focus on V, just 10% of the subjects judged the intended focus to be on S and 12% to be on O.

The percentage of subjects judging the S as being most important was significantly different only by FOCUS [F (3,24) = 154; p<.05], while no effect of MODALITY [F (1,24) = .128; p>.05] nor any interaction [F (3,24) = .301; p>.05] were found. Post-hoc tests revealed significant differences between utterances with narrow focus on the subject and every other focus level. The ANOVA performed on PERCENT-V data revealed again only a main effect of FOCUS [F (3,24) = 98.7; p<.05], and neither a significant difference due to MODALITY [F (1,24) = .803; p>.05] nor a significant interaction [F (3,24) = .93; p>.05]. A post-hoc comparison uncovered significant differences between utterances with focus on V and every other focus level. The data for focus O and broad focus utterances will be discussed in the section below.

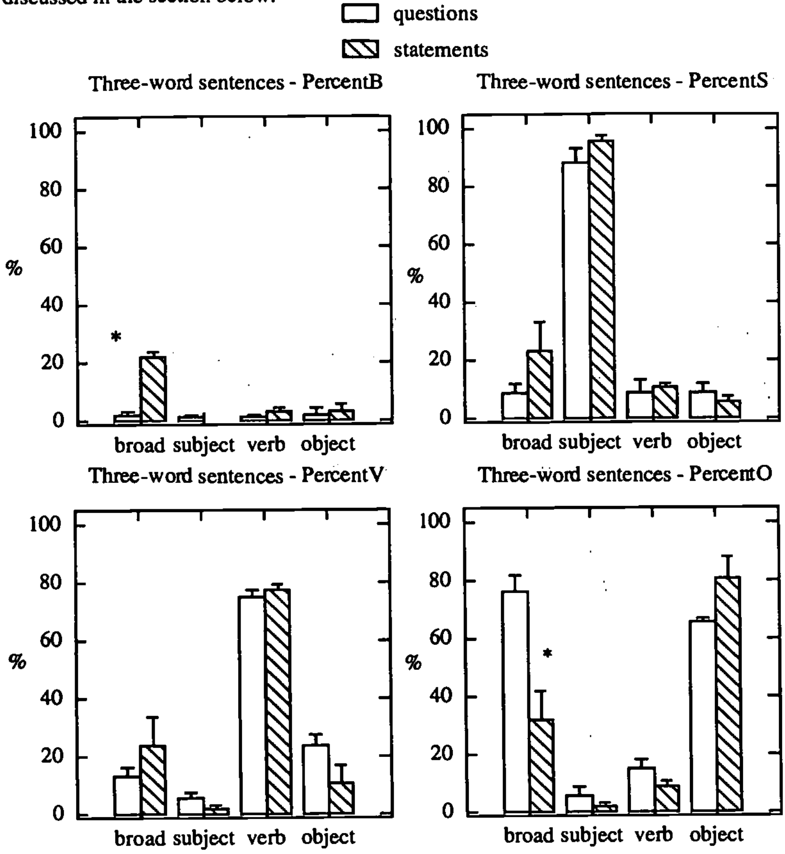


Fig. 8 PERCENT-B (upper left), PERCENT-S (upper right), PERCENT-V (medial, left) and PERCENT-O (medial, right) values by FOCUS (Broad, Subject, Verb) and MODALITY (question vs. statement) for three-word utterances.

Broad focus utterances

For three-word broad focus utterances, the predictions of hypothesis 1 are not completely met. Again, we find that only few (12%) of the responses matched broad focus utterances with questions that elicit narrow focus on O, as opposed to the much higher percentage for utterances with intended narrow focus on O. The prediction of hypothesis 1 is, instead, that these two percentages should be roughly equal, since there is no difference in accent placement between utterances with intended late narrow focus and broad focus.

Upon closer inspection, we notice once again that the effect is due in large part to broad focus statements alone (22%). Indeed, the ANOVA performed on PERCENT-O revealed a significant interaction of the independent variables [F (3,24) = 11.622; p<.05] and a significant effect of MODALITY [F (1,24) = 7.07; p<.05], in addition to the expected effect of FOCUS [F (3,24) = 80.94; p<.05]. A post-hoc test on the focus effect revealed, as in Experiment I, a significant difference between focus B and focus S, focus B and focus V, focus O and focus S and focus O and focus V utterances.

The ANOVA on PERCENT-B revealed a highly significant effect of both MODALITY [F (1,24) = 27.27; p<.05] and FOCUS [F (3,24) = 24.54; p<.05], in addition to a highly significant interaction [F(3,24) = 20.91; p<.05]. Broad focus questions were almost never identified as such, while being associated primarily to questions which are congruent with a focus O question. It is interesting that broad focus questions received higher values of PERCENT-O than questions with narrow focus on O. Figure 9 illustrates the highly ambiguous results for broad focus statements, with values of PERCENT-B, PERCENT-S, PERCENT-V and PERCENT-O juxtaposed. Almost equal values of each of the various response percentages are found for this utterance typology, with PERCENT-O being the highest and PERCENT-B the lowest.

Contrary to previous results, statements with narrow focus on the object presented higher percentage values of focus identified on the object than questions. The data for broad focus statements presented again the greatest variability, at least within the three-word utterance group.

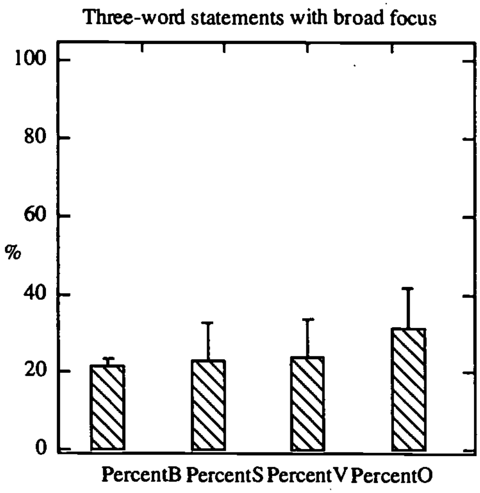


Fig. 9 PERCENT-B, PERCENT-S, PERCENT-V and PERCENT-O for broad focus statements.

GENERAL DISCUSSION

The results of Experiment II remarkably replicate those of Experiment I. For broad focus results in Experiment II, we need to distinguish two-word utterances from three-word utterances. Responses for two-word utterances can be broken down into two groups, in that almost equal PERCENT-V values were found for both broad focus and focus V utterances as a whole, with little difference between questions and statements. These data appeared to support Hypothesis 1 by revealing ambiguity in the identification of breath of focus when accent placement is late. Under this view, broad focus utterances and focus V utterances are both reported by listeners as having intended narrow focus on the last element. The same cannot be said for PERCENT-B results, however. Broad focus utterances produce a much higher value of PERCENT-B than focus V utterances, where the major contribution is due to statements. Also, the value of PERCENT-S for broad focus utterances was almost irrelevant in focus B utterances as a whole, as opposed to the high value reported for the same focus type in Experiment I.

Three-word broad focus utterances presented an overall pattern of responses which is much more comparable to the one found in Experiment I. Three-word statements with focus B produced lower PERCENT-O (which is the latest focus placement in three-word utterances) responses than utterances with focus O. This was not the case for questions, where focus O and focus B utterances patterned identically. Therefore, as a replica of Experiment I, while focus B and focus O questions produced an almost identical pattern of PERCENT-O responses, focus B and focus O statements did not. PERCENT-B values were once again restricted to broad focus statements. Differently from the results of two-word stimuli, broad focus statements produced a considerable amount of responses where a focus S structure is identified.

It is very likely that the question-matching task taps into the notion of focus in a more reliable way than the "assigned importance" task, which is supported by the less variable pattern of overall responses in Experiment II. However, despite the new task, three-word utterances present a picture that highly resembles that of Experiment I. Three-word broad focus statements are almost equally identified as having narrow focus on the subject noun, on the verb, on the object (which obtains slightly higher values relative to the other responses) or as being broadly focused. What is more, the "wide scope" ambiguity appears to be restricted to statements alone, while questions present ambiguity restricted to broad focus and late narrow focus parse.

Overall, the results appear to match our expectations from the outcome of Experiment I, in that we replicated the confusion for broad and late narrow focus questions, as well as the higher uncertainty in the focus structure of broad focus statements. This can be explained with the observed differences in accent type between questions and statements, on one side, as well as between narrow focus and broad focus on the other. As we mentioned above, the "special" status of downstepped accents has been noted for English in the psycholinguistic study of Ayers (1996). Additional evidence for the special status of downstepped pitch accents comes from ambiguity in prosodic transcription: transcribers either disagree in the placement of !H* or fail to recognize its presence in the utterance altogether (Pitrelli et al., 1994; Ross, 1995 cited in Beckman, 1996). Broad focus statements in Neapolitan Italian possess a nuclear accent that has a downstepped quality and is very different from the nuclear accent of narrow focus statements (see Figure 2 above). This could explain the uncertainty on the part of the speakers in focus extraction tasks. The peculiarity of broad focus has also been recently noticed in non-Western languages (see Jin, 1996).

Accent type considerations might predict that the acoustically salient L+H* of Neapolitan Italian questions will facilitate the task of focus extraction for this modality. As we saw from the results of both Experiment I and II, early focus questions, where the nuclear accent is followed by a non-prominent post-nuclear accent, are always correctly identified as having an intended early focus structure. This suggests that a simple positional definition of nuclear stress prominence is not enough and that accent type considerations will have to be formalized to integrate the theory. Such an integrated theory of stress prominence will correctly predict lower ambiguity for accent structures expressing narrow focus scope, because these will be characterized by a phonetically and phonologically more prominent accent type. Similarly, we showed that focus interpretation does not seem to be guided simply by accent placement, but also by choice of accent type. A simplistic theory of the relations between the semantics of focus and the phonology of accent structure that predicts the same parsing difficulty for intended broad focus and late narrow focus will be bound to fail, while an alternative view should predict easier parse for narrow focus structure, where it is relatively simple to retrieve the scope of focus since it is more restricted.

Certainly, the problem of the relation between accent status and prominence on one side and accent structure and focus extraction on the other is particularly intricate for Italian. While it goes with no doubt that Italian possesses some kind of sentence stress, we are still not sure about the constraints operating on it. Among the other things, we still do not know if the sentence stress, or nuclear accent, is, as in English, the pitch-accent that immediately precedes the phrase accent, or if a different characterization is needed. Future research will try to address this issues and will also need to find tasks that will tap into the notion of perceptual prominence in a more reliable way.

CONCLUSION

The present study shows that the predictions of standard generative theories by which broad focus and late narrow focus utterances are confused with each other is true for Neapolitan Italian in the case of questions. However, broad focus statements present a more complex picture, which might be explained by invoking semantic and accent type considerations. On one hand, when focus is broad, i.e. when the focused constituent is larger, the task of parsing focus is more difficult. On the other hand, inherent acoustic prominence due to accent type differences might also play a role, in that the nuclear accent of narrow focus utterances is markedly different from the one that characterizes broad focus utterances.

We also showed that, when appropriate context is presented to the listeners, intended focus can be more reliably identified. The ambiguity of late nuclear accent questions and statements, however, still remains, regardless of the nature of the task. If accent type is, as it seems, an important factor in prominence perception, it will be highly interesting to study what acoustic characteristics are influential in determining more or less prominent accent structures and the interplay of acoustic cues and modality.

REFERENCES

Ayers, G. (1996). *Nuclear Accent Types and Prominence: Some Psycholinguistic Experiments*. OSU Dissertations in Linguistics.

Beckman, M.E. (1986). *Stress and Non-stress Accent.*. Dordrecht, Foris Publications.

Beckman, M.E. (1996). The Parsing of Prosody. *Language and Cognitive Processes*, 11, pp. 17-67.

Beckman, M.E. & Ayers, G. (1994). *Guidelines for ToBI Labeling*, vers. 2.0, Ms., Ohio State University.

Bruce, G. (1982). Developing the Swedish intonation model. *Working Papers* 22, Department of Linguistics and Phonetics, Lund University, pp. 51-116.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. New York, Harper & Row.

Chomsky, N. (1971). Deep Structure, Surface Structure, and Semantic Interpretation. D. Steinberg and L. Jakobovits (eds.), *Semantics: an Interdisciplinary Reader in Philosophy, Linguistics and Psychology*. Cambridge, CUP, pp. 183-216.

Crystal, D. (1969). *Prosodic systems and intonation in English.* Cambridge, CUP.

Fry, D.B. (1958). Experiments in the perception of stress. *Language and Speech*, 1, pp. 126-152

Halliday, M.A.K (1967). *Intonation and grammar in British English.* The Hague, Mouton.

Horne, M. (1991). Phonetic correlates of the new/given parameter. *Proceedings of ICPhS '91*, Aix-en Provence, 5, pp. 230-233.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar.* Cambridge, MA, MIT Press.

Jin, S. (1996). *An Acoustic Study of Sentence Stress in Mandarin Chinese.* Ph.D. dissertation, OSU.

Ladd, R.D (1980). *The structure of intonational meaning.* Bloomington, Indiana University Press.

Ladd, R.D (1996). *Intonational Phonology.* Cambridge, CUP.

Lehiste, I. (1970). *Suprasegmentals.* MIT, Cambridge, MA.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8, pp. 249-336.

Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *JASA*, 66, pp. 363-9.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation.* Unpublished Ph.D. thesis, MIT.

Pierrehumbert, J. (1993). Alignment and Prosodic Heads. A. Kathol and M. Bernstein (eds.), *ESCOL '93*, pp. 268-286.

Selkirk, E. (1984). *Phonology and Syntax: The Relation between Sound and Structure.* Cambridge, MA, MIT Press.

The Northern Cities Shift in the Heartland? A Study of Radio Speech in Columbus, Ohio*

Steve Hartman Keiser, Frans Hinskens, Bettina Migge, and Elizabeth A. Strand[1]
shkeiser@ling.ohio-state.edu
F.Hinskens@let.kun.nl
bmigge@ling.ohio-state.edu
estrand@ling.ohio-state.edu

Abstract: Variation in vowel height and diphthongal/monophthongal character of the vowels /æ/ and /a/ are studied in the speech of two speakers from central Ohio in order to measure their participation in the sequence of vowel system changes commonly referred to as the Northern Cities Shift (Labov, 1994). The data were gathered from radio shows for which the speakers served as announcers. Determinations of vowel height and diphthongal nature of vowels were made by auditory judgment of the researchers and were correlated with acoustic measurements of *F1* and *F2* frequencies. The results suggest that the vowel system of the central Ohio dialect is undergoing change, but are inconclusive as to whether this change indicates participation in the Northern Cities Shift. Detailed analyses of social and linguistic factors correlated with the tensing and raising of /æ/ are offered.

INTRODUCTION

Recent research in American dialectology has focused on differences in vowel systems as a means for differentiating regional dialects (e.g., Labov, 1991). It has been demonstrated, for example, that the dialects of the southern states can be characterized by a raising and tensing of front, lax vowels and a concomitant lowering and laxing of front, tense vowels along with a somewhat fronted production of back vowels. This collection of

* We want to thank Keith Johnson for helping guide us through the intricacies of the acoustic analyses involved in this project, and for his timely and generous technical support. We offer a heartfelt thanks to Rebecca Herman for giving us hours of her time in listening to our data. Thanks also to Mike Cahill for rating our data. Additionally, we are grateful to Norma Mendoza-Denton for her helpful insights and comments on earlier drafts of this paper. We are indebted to Jignesh Patel, who went beyond the call of duty to help us recover our virus-obliterated computer files. And finally, a special thanks to "Red" and "Daniel" for introducing us to their radio stations and for sharing their speech across the airwaves.

[1] Authors are listed in alphabetical order. Steve Hartman Keiser, Bettina Migge, and Liz Strand are affiliated with the Department of Linguistics at The Ohio State University, and Frans Hinskens is affiliated with the Department of General Linguistics and Dialectology, University of Nijmegen, the Netherlands. This project is the result of collaboration that occurred while Frans Hinskens was a visiting professor in Linguistics at The Ohio State University in 1996-97.

shifts in vowel quality with respect to Standard American English (SAE) has been dubbed the Southern Vowel Shift (see Figure 1).

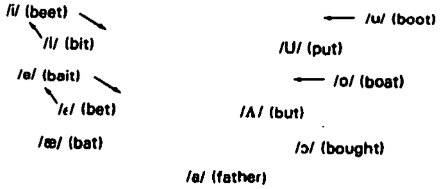


Figure 1. Vowel rotation in the Southern Vowel Shift (adapted from Wolfram, 1991:87).

A number of recent studies have focused on a separate shift in the vowel systems of dialects spoken in certain urban areas of the northern United States, including Rochester, New York; Buffalo; Detroit; and Chicago, among others (Eckert & McConnell-Ginet, 1995; Gordon, 1996; Labov, 1991, 1996). Many Euro-American English speakers in these cities produce short, or tense, vowels that are markedly different from their SAE counterparts: /æ/ is raised and tensed, /a/ is fronted, /ɔ/ is lowered and slightly fronted, /ɪ/ is lowered, /ɛ/ is produced more centrally, and /ʌ/ is backed (as is shown in Figure 2). This set of shifts is generally referred to as the Northern Cities Chain Shift, or Northern Cities Shift (NCS).

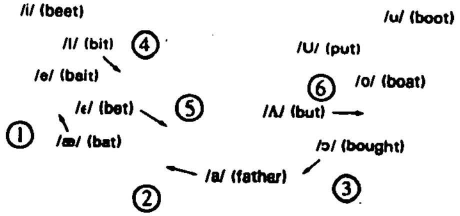


Figure 2. Vowel rotation in the Northern Cities Shift (adapted from Wolfram, 1991:87). Circled numbers are added to indicate the ordering of vowel movements, with "1" occurring first and the others following in order (Labov, 1994:195).[2]

These movements in the vowel space are not random. The vowel space appears to act as a connected system. As speakers' production of a given vowel shifts, the shifted vowel both impinges on the phonetic space occupied by other vowels and leaves behind a void in the phonetic space for another vowel to move into. This can result in the collapsing

[2] The ordering of the final three changes has not been conclusively determined. Labov (1996), e.g., posits that /ɛ/→/ʌ/ is "4," /ʌ/→/ɔ/ is "5," and the centralization of /ɪ/ is "6."

of what were once two vowel phonemes into one, or it can "push" or "pull" adjacent vowels to move as though they were linked by a chain; thus, the term "chain shift." Hock & Joseph (1996:134) note that chain shifts are conditioned not only by purely phonetic factors, but are responsive to abstract cognitive structure as well. This seems plausible for any but the initial shift.

The Northern Cities Shift has been touted as "a massive change that bears no resemblance to any chain shift previously recorded in the history of the [English] language" (Labov, 1994:10). What is more, it is a chain shift in progress across temporal and geographical space, as demonstrated by the fact that some speakers (often younger urbanites) produce shifted variants for all of the vowels in question, while other speakers (often older ruralites) shift only some of the vowels. Further studies indicate that the shift is spreading to surrounding rural areas in the north (Gordon, 1996; Ito, 1996).

Although the northern Ohio cities of Cleveland and Toledo participate in the NCS, no such chain shift phenomenon is claimed to be diagnostic of the dialects of central Ohioans[3]. In most dialectology maps, the region surrounding the capital, Columbus, is classified as "lower north" or "midland" (see, e.g., Wolfram, 1991:83-5). However, Ohio as a whole, and central Ohio in particular, have been largely ignored in dialect studies, leaving a gap in our knowledge about the dialect(s) of the millions of speakers who inhabit this area. For these reasons, Labov has characterized this area as "the mysterious midlands" (Labov, p.c.).

The goal of this paper is to take a small step toward closing this knowledge gap. Specifically, given the apparent vitality of the NCS and our position as researchers located in a major urban area[4] adjacent to the traditional northern dialect region, we raise the following questions: to what extent is the NCS making inroads in the speech of central Ohio speakers? And to what extent are individual vowel shifts in central Ohio conditioned by internal and external factors? This paper is a preliminary report of this work in progress.

METHODS

In this pilot study, we limit the number of dependent variables to just two of a possible six vowel shifts in the NCS, and utilize an uncommon method of data collection: recording radio speech. As will be discussed below, due to the small subject pool and limited stylistic variation in the data, certain aspects of our socio-stylistic analysis can only be considered preliminary.

Included in this section are the following subsections: "Selection of Dependent Variables," "The Corpus: Radio Speech," "Selection of Speakers," "Data Collection," "Analysis for the Tensing and Raising of /æ/: Scale Development and Scale Validation," "Vowel Spaces and the Interaction of Height and Tenseness of /æ/," "Selection of Indpendent Variables for the Study of /æ/ Variation," and finally, "Statistical Methods."

Selection of Dependent Variables

Of the six different vowel shifts which together make up the NCS, the present project is confined to only two shifts (our dependent variables): the tensing and raising of /æ/, and the fronting of /a/.

[3] It should be noted that the term "central Ohio" can be used to refer to a rather large area of several dozen counties with a heterogeneous populace. Given this imprecise definition of "central Ohio," it is possible to say that certain aspects of the Southern Shift are present in the speech of some central Ohioans (e.g., *fish* as [fiʃ], but possibly conditioned by the following palatal segment).

[4] The population of the greater Columbus metropolitan area is 1.35 million, according to the 1990 census.

We would expect speakers from Columbus—an urban area which neighbors the northern dialect region—to either exhibit none of the NCS shifts, or only the initial shifts in the chain. This requires an investigation into the relative chronology of the NCS.

A temporal ordering of the vowel shifts in the NCS places the raising of /æ/ as the first shift to take place (the circled numbers in Figure 2 indicate the order in which the vowel movements are taking place). This shift appears to be nearing completion in many communities where the NCS is active. Selecting /æ/ as the sole dependent variable, however, would not provide conclusive evidence of the NCS, given the volatile nature of /æ/ in American English. For example, the raising of /æ/ can also be found in many southern varieties of American English.

Therefore, a second shift, the fronting of /a/, is examined in conjunction with the raising of /æ/. While the raising of /æ/ is not unique to the NCS, the fronting of /a/ is.[5] This shift is chronologically second in the chain, and is considered to be a change in mid-course. These two shifts can be seen as functioning as a pull or drag chain, with the movement of /æ/ clearing the phonetic space into which /a/ can then move (Labov, 1994:195). Alternatively, we might wonder whether this second shift is feeding into the first, so that tokens of /a/ are shifted to /æ/ and then raised, resulting in the elimination of all low vowels. The fact that the tensing and raising of /æ/ is nearing completion in northern cities, whereas the fronting of /a/ is a much more recent phenomenon, makes this an unlikely scenario.

We selected these two vowels, /æ/ and /a/, involved in the initial shifts in the NCS, as our dependent variables. Taken together, these two vowels should provide a clear indication of the presence or absence of NCS-like shifting in the speech of central Ohioans.

The Corpus: Radio Speech

We wanted a reasonable way to record large samples of speech for phonetic analysis. We decided to gather data from local radio on-air personalities, a method which has been successfully employed by other researchers (Bell, 1984; Van de Velde & van Hout, 1996; Van de Velde, et al., 1996).

First, we identified two radio stations which can be reasonably claimed to differ in the demographics of their listener base. One station plays an all-classical music format and the other plays exclusively contemporary country music. Van de Velde, et al. (1996:5) state that in the selection of stations for such research, one should look for stations targeting clearly different and relatively single-layered audiences. Our idea in taking this step is that the speech of announcers on the two stations will, to some extent, reflect the speech of the audience that they target. The theoretical framework supporting this claim is Bell's work on speech style as audience design, which assumes that "persons respond mainly to other persons, that speakers take most account of hearers in designing their talk" (Bell, 1984:159).[6] There are four audience roles in Bell's theory which differ in the degree of influence that they exert on a speaker's style (listed in descending order of influence): addressee, auditor, overhearer, and eavesdropper. In radio speech, the addressees are the station's target audience. Auditors, overhearers, and eavesdroppers make up the rest of the potential listening public (177). Here we are most concerned with the audience which has

[5] Whereas the Northern Cities dialects front /a/, certain New York City dialects raise /a/ to /o^{ə}/ or even /u^{ə}/ (Hock & Joseph, 1996:134); in the Southern Vowel Shift, /a/ is entirely stable.

[6] Labov's study of post-vocalic [r] in the speech of New York City department store personnel (1972) is an example of an application of a method similar to ours, but which predates Bell's theoretical framework. Labov identified three department stores serving clientele of demonstrably different socioeconomic statuses. The speech of clerks in each store was then interpreted as reflecting the speech of the SES of the customers they served.

the most impact on speaker style: the addressees (target audience) for each station. In short, these radio stations, and therefore the announcers employed by them (and their speech), can be considered reflections of the targeted audience and their respective socioeconomic strata in society.

One indication of the difference in listener demographics of the two stations included in our study came from a personal communication with the station manager of the classical station. In terms of a profile for the station's prototypical listener/supporter, he submitted the following description: over age 35, affluent, well-educated, "National Public Radio-type" person. This appears to fit Van de Velde's single-layer audience criterion. It is less clear how or whether the country station audience is similarly "single-layered." While we cannot conclude that the listeners of the country music station are categorically younger, poorer, and less educated than the classical station listeners, it is reasonable to assume that the classical station prototype would describe only a small subset of the listener base of the country station. We would expect, then, that on average, measures of socioeconomic status would show a difference between the audiences of the two radio stations: the classical station would represent a higher socioeconomic status, and the country station would represent a lower socioeconomic status.

A second criterion in selecting the stations was a control variable, this being that both stations aim at the same greater Columbus metropolitan area audience. Two pieces of evidence show that this "locality constraint" is satisfied with our selection of these two stations: the classical music station has an hourly Columbus traffic update, and the country station has many listeners who call in to the on-air programs and identify themselves as residents of the greater Columbus metropolitan area.

Selection of Speakers

We identified two native speakers of Central Ohio English, both of whom were radio announcers on Columbus-area stations. The speaker from the country station was an approximately 25-year-old male native of Columbus whom we will call "Red." The speaker from the classical station was a male native of Newark (30 miles east of Columbus) in his mid-forties, whom we will refer to as "Daniel." Sex of the speakers was treated as a control variable. Potential influences due to age difference are not considered here.

Data Collection

For both speakers, speech samples were recorded from the radio during their on-air shows. "Red's" show aired in the afternoon and evening and consisted of him announcing the songs to be played, as well as taking a number of calls from listeners requesting favorite songs or taking part in on-air promotions. "Daniel" was recorded during the classical station's quarterly fund-raising drive, and his participation consisted of a number of monologues touting the benefits of supporting his station, as well as some interaction with an on-air colleague.

For both speakers, two "style" levels were recorded, what we termed "monologue" and "dialogue." We used "monologue" to refer to relatively scripted talk (almost always addressed to the listening audience), and "dialogue" to refer to talk that occurred either between the announcer and listeners calling in to the station, or between the announcer and his on-air colleagues. This style variable is included to test one prediction of Bell's theory. According to Bell's design, "the mass auditors [the targeted listening audience] are likely to be more important to a communicator than the immediate addressees [the on-air collegues or listeners calling in to the station]" (Bell, 1984:177). In other words, all radio speech is addressed to the mass auditors, and so therefore there should not be a significant style difference between the monologue and dialogue situations. The opposite hypothesis is also defensible: that in dialogue, the amount of attention paid to speech will be less than that in announcement-types of speech (monologue), yielding a style difference.[7]

Analysis of Tensing and Raising of /æ/: Scale Development and Scale Validation

We made a preliminary auditory analysis of tokens of /æ/ and /a/ in the recorded data from our two speakers. This analysis revealed a high degree of variablity in the tokens of /æ/ and very little in the tokens of /a/. For this reason, we leave investigation of the possible variation in /a/ to a future study, and rather focus on the development of more refined analyses for the examination of the tensing and raising of /æ/.

We initially conducted quantitative analyses of 343 realizations of morphemes containing the segment /æ/. These analyses consisted of several steps. First, in order to code each token for its level of tensing and raising, we needed to develop scales to describe each of these characteristics. We did this by first making narrow phonetic transcriptions of 30 token types. We carefully grouped these 30 original types into several categories which allowed us to create two descriptors specifically for the purpose of describing the data in this study, a height scale and a tenseness qualifier. The former was a five-point scale for which the points were salient vowel height categories for each of the author-listeners (as is shown in Figure 3 below).

Height rating	Phonetic symbol[8]	Example word
5	ɪ	*bit*
4	e	*bait*
3	ɛ	*bet*
2	Æ[9]	
1	æ	*bat*

Figure 3. Height scale developed from our Columbus data.

We denoted the tensing (diphthongal/monophthongal) quality of the vowel as shown in Figure 4:

[7] In the Labovian definition of style, the relative amount of attention paid to speech plays a central role in the (relatively formal or informal) style of the speech (Labov, 1972:Chapt. 3).

[8] These "phonetic symbols" are generally equivalent to their IPA counterparts, with the exception of Æ, discussed in Footnote 9 below.

[9] This symbol, "Æ," which we refer to as "capital ash," represents a salient category between lower /æ/ and higher /ɛ/ for this variety. For example, the token vowel in question was compared with the vowel in each member of the minimal pair *bat* [bæt] and *bet* [bɛt]. When the vowel was perceived as neither of these—i.e., higher than the vowel in *bat* but lower than the vowel in *bet*—it was classified as /Æ/. While salient to the authors and data coders, this category is not contrastive, thus does not exist phonemically in Standard American English.

Diph. rating	Vowel status
+	"diphthongal"
-	"monophthongal"

Figure 4. Means of denoting diphthongization in our Columbus data.

Using these two scales, we scored 131 morphemes containing /æ/ for the country station speaker, and 212 morphemes containing /æ/ for the classical station speaker. Three of the authors individually scored each token (yielding the individual scores refered to later in the paper), and then scored each token together as a group (yielding the group scores). It is the group scores that form the basis for our statistical analyses, and will henceforth be refered to as the *group auditory judgments*. In creating a data matrix from the coded tokens, we included information for each of the internal, independent variables (as described below).

We validated our auditory scales in three ways: (1) comparisons of the group auditory judgements to those of an independent listener; (2) factor analysis and crosstabulations of the individual ratings of the three authors; and (3) correlation of the group auditory judgments with formant measurements.

As the first means of validating our scales, a fourth listener independently scored each realization. We then compared the group auditory judgments for the height scale with the scores of this listener, using a Spearman rank correlation. Results show that for N=343, the correlation coefficient is .4707, with significance at .01. This shows a significant but relatively weak positive correlation between the independent listener's judgments and the group auditory judgments.

In order to compare our judgments on diphthongal/monophthongal character, we crosstabulated the group auditory judgments by the independent listener's judgments. The results indicate 94.4% mutual agreement between the author group and the independent listener for scoring monophthongs,[10] but a 34.2% mutual agreement for scoring diphthongs.[11] Results show a Cramer's V value of .36588 (which gives an indication of the degree of association). Overall, there is a relatively high level of mutual agreement, since the author group and the independent listener agree on the monophthongal/diphthongal nature of 81.6% of the tokens (280 out of a possible 343 realizations).

As a second means of validation, here for our auditory height scale, we conducted factor analyses of the individual ratings of the three authors. Only one factor was extracted. The factor loadings are relatively high, ranging between .85 (Bettina) and .89 (Liz). This indicates that the judgments of the three authors are placed on the same dimension.[12] In other words, the three authors are all utilizing the scale in the same way, which validates the psychological reality of the height scale.

[10] The group auditory judgment categorized 270 tokens as monophthongs out of a possible 343. Out of these 270 tokens, the independent listener agreed that 255 (or, 94.4%) were monophthongs.

[11] The group auditory judgment categorized 73 tokens as diphthongs out of a possible 343. Out of these 73 tokens, the independent listener agreed that 25 (or, 34.2%) were diphthongs. This lower agreement rate for diphthongs may stem from varying notions of what constitutes a diphthong.

[12] However plausible the outcomes of the factor analysis, reliability analyses would be preferred for this step in the validation. We have not yet run these analyses.

In examining our means of rating diphthongization, we crosstabulated the judgments of the individual authors on a pair-wise basis: Bettina vs. Liz, Bettina vs. Steve, and Liz vs. Steve, as shown in Table 1 below.

CELLS:		LIZ			STEVE		
Raw scores		diph.	monoph.	TOTAL	diph.	monoph.	TOTAL
BETTINA	diph.	25	14	**39**	26	13	**39**
	monoph.	71	233	**304**	50	254	**304**
	TOTAL	**96**	**247**	**343**	**76**	**267**	**343**
LIZ	diph.				49	47	**96**
	monoph.		X		27	220	**247**
	TOTAL				**76**	**267**	**343**

Table 1. Results of crosstabulations for diphthongal and monophthongal agreements between the three individual authors, given in raw scores.[13] Results are significant at a level of .01 using a measure of Pearson chi-square probability.

There is strong general agreement between each of the authors regarding the diphthongal/monophthongal nature of the tokens. The highest level of agreement is between Steve and Liz (Cramer's V=.43359, significant at the .01 level from a Pearson chi-square probability). The lowest level of agreement is between Bettina and Liz (Cramer's V=.28813, significant at .01).[14]

As a third means of validation, here involving our tenseness qualifier, we took formant measurements of the *F1* (first formant) and *F2* (second formant) values for each occurrence of /æ/. Using the Kay Elemetrics Computerized Speech Laboratory (CSL) system, each vowel token was first digitized (using a 10 kHz sampling rate) and spectrograms and LPC ("linear predictive coding") formant tracks were then generated, from which the formants were measured. The formants were measured by visual inspection of the spectrograms and LPC formant tracks at three points in each vowel: at or near vowel onset, at a point in the middle of the vowel, and at or near the end of the vowel.

[13] For example, in the first column under diphthongal scores for "Liz" crossed with "Bettina," the number "25" indicates the number of tokens that Liz rated as "diphthongal" which Bettina also rated as "diphthongal." "71" indicates the number of tokens that Liz rated as "diphthongal" which Bettina rated as "monophthongal." And finally, "96" is the total number of tokens rated as "diphthongal" by Liz (out of 343 total.)

[14] The fact that we find the strongest similarity between the authors Liz and Steve is plausible, since they are both L1 speakers of neighboring dialects of American English. The fact that the similarity between Bettina and the other authors is lower is also plausible, since Bettina is an L2 speaker of English and therefore might be using different strategies in the categorization of tokens as monophthongs or diphthongs. These findings seem to provide indirect indication of the reliability of our tenseness qualifier.

The LPC parameters were 14 coefficients, Hamming windowing, and 0.97 preemphasis. (See Figure 5 for a representation of these formant measurements.)

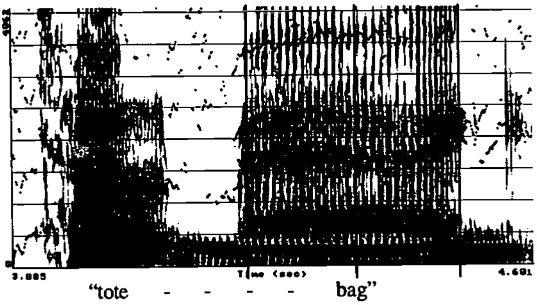


Figure 5. Representation of the measurement of *F1* and *F2* values at three points in the /æ/ token. "X's" are placed approximately at locations where the formant measurements were taken from points on the LPC formant track (which is shown as a light-grey trace throughout each formant). (The word represented here is *totebag*, as spoken by "Daniel.")

Vertical bars in the spectrogram in Figure 5 represent the the initial, middle, and end points within each vowel token at which measurements of *F1* and *F2* were taken. These formant measurements were included in our data matrix.

The auditory judgments were then correlated with the mid-values of *F1* and *F2*, as well as the intitial values of *F1* and *F2*, to verify height.[15] Results of correlations indicate that, as expected, height is negatively correlated with initial *F1* (correlation coefficient = -.4776) and mid *F1* (-.2614), and positively correlated with initial *F2* (.4372) and mid *F2* (.5963). Results are significant at a level of .01.

To verify the auditory judgments of the diphthongal nature of the tokens, the judgment scores were correlated with the difference in initial and final measurements (initial *F1* - final *F1*; inital *F2* - final *F2*) and with vowel duration. Results of these correlations are shown in Table 2 below.

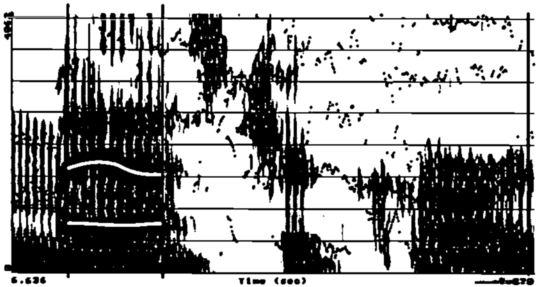
	MEAN		SD		Mean			
	Diph.	Monoph.	Diph.	Monoph.	diff.	t	df	Signif.
F1 diff.	-48 Hz	-22 Hz	126 Hz	289 Hz	26 Hz	.74	336	.461
F2 diff.	169 Hz	33 Hz	365 Hz	282 Hz	137 Hz	-3.43	339	.001
duration	126 ms	105 ms	82 ms	78 ms	22 ms	-2.10	337	.037

Table 2. Results of t-tests for the effect of *F1* and *F2* differences and duration on auditory judgments of diphthongization.

[15] We tested our validated judgments of dependent variables against the independent variables using ANOVA and logistic regression in the SPSS software package, Windows version 6.1.3.

As the results in Table 2 indicate, "*F2* difference" and "duration" are significant. These acoustic measures corroborate our judgments of tensing: the greater the change in *F2* and the greater the duration of the token, the more likely we are to judge it as a diphthong.

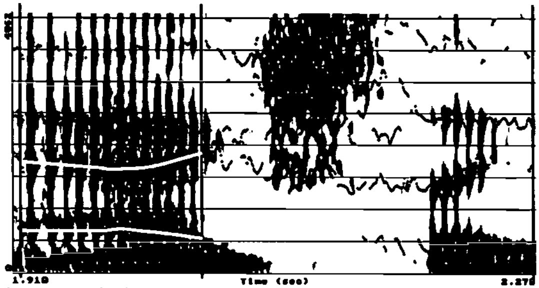
Spectrograms of /æ/ tokens of various heights and tenseness are included in the six figures below to further illustrate the validity of our auditory judgments. Examples from the speech of "Daniel" are given in the first three figures, and examples from the speech of "Red" are given in next three figures. The relevant /æ/ token in each spectrogram is segmented with vertical lines, and white traces are added (following the original LPC traces) to track *F1* and *F2* in these tokens. A loose phonetic transcription and /æ/ token score are provided directly beneath each spectrogram.



[m æ s tʰ ɚ kʰ a ɹd]
[1-]

Figure 6. Spectrogram of "Daniel" saying *Mastercard*, with /æ/ token rated [1-].

Figure 6, Figure 7, and Figure 8 give spectrograms of tokens produced by "Daniel," our classical station announcer. For "Daniel's" production of the word *Mastercard*, represented in Figure 6, the /æ/ token was scored as [1-], meaning that it received a height score of [1] (approximating /æ/, see Figure 3) and a tensing evaluation of [-], meaning that it was perceived as being monophthongal. Acoustically, the relatively high *F1* and reasonably level formants throughout the vowel give credence to this score.



[ʔÆ k ʃ l iʸ]
[2-]

Figure 7. Spectrogram of "Daniel" saying *actually*, with /æ/ token rated [2-].

Figure 7 is a representation of "Daniel's" production of *actually*, in which the /æ/ token was rated as [2-] by the group. *F1* is slightly lower and *F2* is slightly higher than the formants in the /æ/ token in Figure 6, which is consistent with the higher-rated vowel shown here in Figure 7. Considering that upward movement of *F2* near the end of the vowel is related to coarticulation with the following velar segment (thus the characteristic "velar pinch" of *F2* and *F3*), we might say that formant movement is still consistent with a segment perceived to be monophthongal.

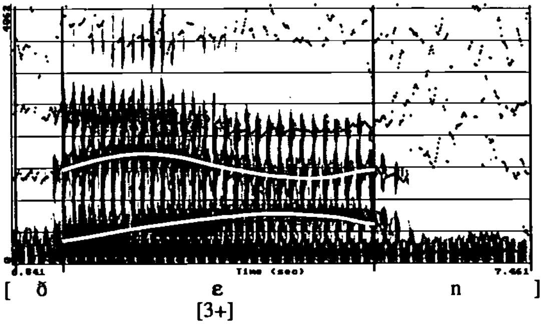


Figure 8. Spectrogram of "Daniel" saying *than*, with /æ/ token rated [3+].

In Figure 8, we see a spectrogram of "Daniel's" production of *than*, which was scored as [3+], meaning that this vowel token was perceived as higher than the vowel in Figure 7, as well as diphthongal. At least from vowel-onset to near the midpoint of the vowel, *F1* is lower than it is in the vowel tokens in the figures above, indicating increased height. The formant movements that are evident in Figure 8 are most likely instrumental in the vowel being perceived as diphthongal.

Figure 9, Figure 10, and Figure 11 give spectrograms of tokens produced by "Red," our country station announcer. One thing to note (and disregard) in these figures are the background music formants that run faintly throughout the spectrograms, which result from "Red's" talking over music during most of his radio show.

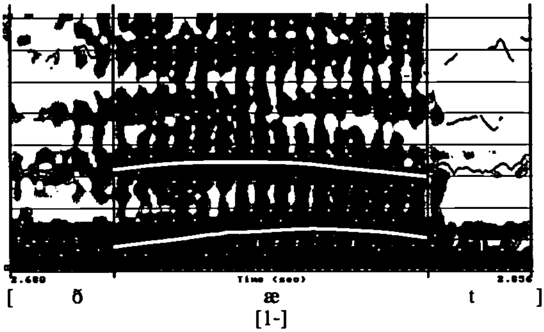


Figure 9. Spectrogram of "Red" saying *that*, with /æ/ token rated [1-].

In Figure 9, "Red's" production of *that* illustrates a vowel token that was scored as [1-]. The formants are relatively stable throughout the vowel, indicating what was perceived as a monophthongal quality, despite the fact that *F1* may seem a bit low to have rendered a score of [1] (indicating no perceived raising of the /æ/ token).

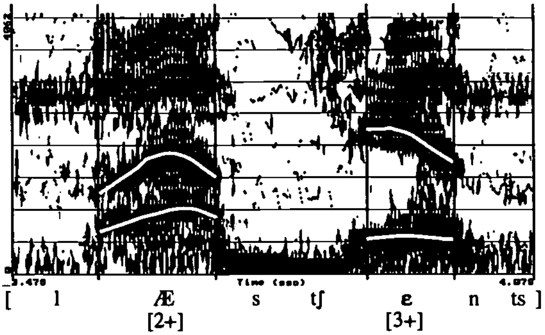


Figure 10. Spectrogram of "Red" saying *last chance*, with the first /æ/ token rated [2+] and the second /æ/ token rated [3+].

Figure 10 offers a good comparison of two /æ/ tokens in different contexts within the same utterance. Both tokens are rated as diphthongal, but the first token (in the word *last*) is scored as [2], one height-level lower than the next token (in the word *chance*), which was scored [3]. *F1* and *F2* in these two tokens seem to vary accordingly: *F1* is higher and *F2* is lower in the vowel that is perceived as being lower, and *F1* is lower with *F2* higher in the vowel that is perceived as higher. This figure also illustrates the tendancy for previous liquids to restrict /æ/-raising (as in *last*) and following nasals to positively affect /æ/-raising (as in *chance*), as will be discussed in the Findings section later in the paper.

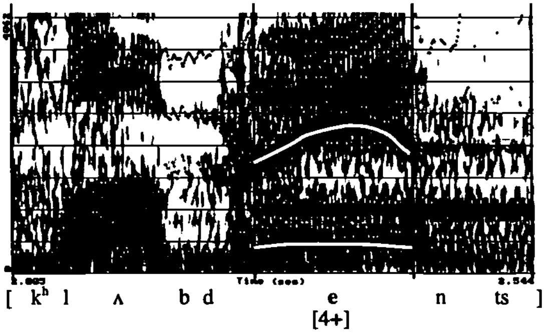


Figure 11. Spectrogram of "Red" saying *Club Dance*, with /æ/ token rated [4+].

Finally, Figure 11 shows a spectrogram of "Red's" production of the term *Club*

Dance, a frequent utterance during his promotions of night club activity in Columbus. The high realization of this token, scored [4+], is typical of /æ/ tokens followed by nasal segments in "Red's" speech.

Vowel Spaces and the Interaction of Height and Tenseness of /æ/

The distributions of tokens of /æ/ for each of the speakers are shown in Figure 12 and Figure 13 below. The general vowel space is delimited by anchor tokens of /i/ and /u/, with the addition of the rated /æ/ tokens. Individual tokens of /æ/ are marked on the figures using the phonetic symbols which we use to represent each vowel height in our scale, as shown in Figure 3. An exception to this is our usage of the symbol "a" in the figures below to represent "Æ" *capital ash* of Figure 3, due to limitations in our plotting software. Ellipses enclose 95% of the data points for each vowel height category. These ellipses are the result of a principal components analysis of variation for each vowel height category, on which two standard deviations are then calculated.[16] It should be clear from these figures that raising takes place in the speech of both Red and Daniel, but there is greater range of variation in the height of Red's /æ/ tokens. Red's *F1* measurements vary from about 400 Hz to 1000 Hz, while Daniel's measurements vary from about 300 Hz to 800 Hz. In terms of *F2*, Red shows variation from 1300 Hz to 2700 Hz, and Daniel shows variation from 1300 Hz to 2100 Hz.

The distributions of tokens of /æ/ for each of the speakers, as well as sample representations of /i/ and /u/ from the data, are shown in Figure 12 and Figure 13 on the following page.[17]

[16] The "principal components" of the data clouds for each vowel height category are two regression "lines" through each cloud at 90° angles to each other. These right-angle lines then become the axes for which the standard deviations are found for the category.

[17] Interesting to note in these vowel spaces (though not relevant to the present study) is the greater acoustic variability of /u/ than /i/, especially evident in the representative vowel space of "Red." In general, the greater front-back "displacement" of /u/ is due in part to contextual variation; a following front consonant will cause /u/ to be fronted, while a following back consonant will cause /u/ to be backed, effects which are evident in the acoustics of the vowel in these different environments. The vowel /i/, however, is much more resistant to contextual influence since it is produced with a high degree of tongue bracing against the walls of the mouth (Fujimura & Kakita, 1975). This bracing tends to stabilize /i/ against the level of perturbation due to contextual influence which can be seen in /u/.

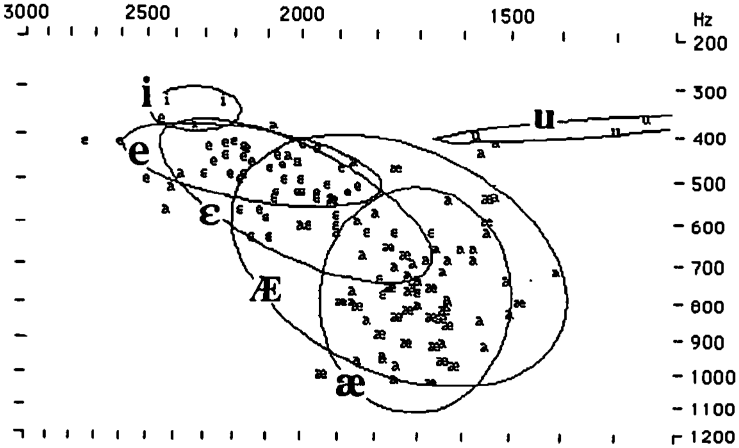


Figure 12. Vowel space of country station announcer "Red."

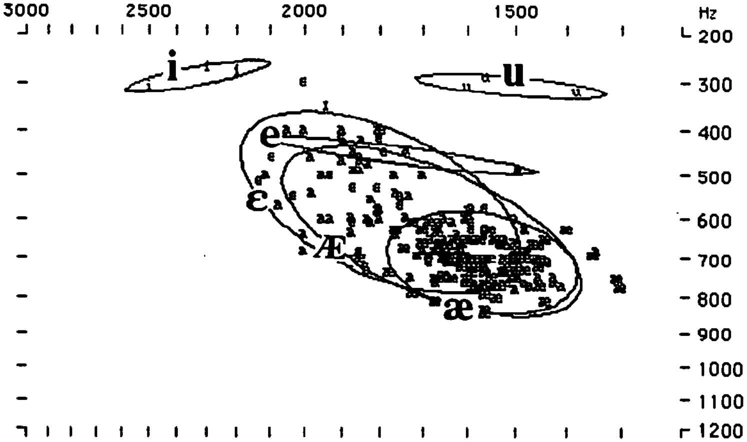


Figure 13. Vowel space of classical station announcer "Daniel."[18]

[18] Note the single token of [I] that is plotted in "Daniel's" vowel space, with an *F1* of approximately 350 Hz, and an *F2* of approximately 1950 Hz. This /æ/ token was produced in the word *thank*. Since this was the only token in our data to be rated [I] (i.e., scored [5] according to our height scale in Figure 3), it is not included in an ellipse.

Analysis of our two dependent variables height and tenseness (the latter indicated by diphthongization) show that they interact. As vowel height increases, the percentage of diphthongs increases for each height category. That is, the more a token of /æ/ is raised, the more likely it is to also be tensed. (See Table 3 below.) So, for example, only 10.9% of unraised tokens of /æ/ are diphthongized, whereas 19.1% of tokens raised to the next height ([Æ]) are diphthongized, 22.2% of tokens raised to [ɛ] are diphthongized, and finally, 76.9% of tokens raised to [e] are diphthongized.

		DIPHTONS		
	Count Row Pct Col Pct	diphthon giz d	monophth ong m	Row Total
HEIGHT				
ae	1	12 10.9 16.4	98 89.1 36.3	110 32.1
AE	2	29 19.1 39.7	123 80.9 45.6	152 44.3
E	3	12 22.2 16.4	42 77.8 15.6	54 15.7
e	4	20 76.9 27.4	6 23.1 2.2	26 7.6
I	5		1 100.0 .4	1 .3
	Column Total	73 21.3	270 78.7	343 100.0

Table 3. Crosstabulation of Height by Diphthongization / Monophthongization for group auditory judgments of both speakers' data.

These data suggest that tensing and raising of /æ/ are not independent phenomena in central Ohio (but see the discussions of the factor groups "proximity to right-hand word boundary" and "stress" below).

Selection of Independent Variables for the Study of /æ/ Variation

A preliminary auditory analysis of tokens of /æ/ and /a/ in the recorded data from our two speakers revealed a high degree of variablity in the tokens of /æ/ and very little in the tokens of /a/. For this reason, we focused our study to an examination of /æ/ raising and tensing phenomena. We leave investigation of the possible variation in /a/ to a future study.

Since our study includes speech data from only two individuals within an audience design framework, we do not factor in social characteristics of the individuals. We selected only two external (social) independent variables to investigate. (These factors were introduced above in the subsections "The Corpus: Radio Speech" and "Data Collection," and are briefly discussed below.)

EXTERNAL FACTOR GROUPS FOR THE TENSING AND RAISING OF /æ/

1. socioeconomic status (SES), defined by station
   - classical station (reflecting a relatively higher average SES with respect to the country station)
   - country station (reflecting a relatively lower average SES with respect to the classical station)
2. style
   - monologue
   - dialogue

A review of previous research on similar vowel shifts led us to select the following internal (linguistic) independent variables for our study of /æ/ (see discussion of each of these variables below):

INTERNAL FACTOR GROUPS FOR THE TENSING AND RAISING OF /æ/

1. membership in *mad, bad, glad* lexical class
   - belong to class
   - end in *-ad* but are not *mad, bad, glad*
   - other
2. grammatical category
   - preterite strong verb
   - preterite irregular verb
   - preterite regular (weak) verb
   - other verbs (non-preterite)
   - non-verb
3. right-hand morphological boundary
   - word
   - Class 1 suffix[19]
   - Class 2 suffix
   - inflectional suffix
4. proximity to right-hand word boundary, measured in terms of syllables
5. stress
   - stressed monosyllabic word
   - primary
   - secondary
6. preceding phonetic segment(s) (in the case of a morpheme-internal cluster, all segments of the cluster were noted)
7. following phonetic segment
8. syllable membership of the following consonant
   - following consonant(s) in the same syllable ("tautosyllabic")
   - following consonant shared with the next syllable ("ambisyllabic")[20]

[19] From O'Grady, et al. (1997:132-3): "it is common to distinguish between two types of derivational affixes in English. Class 1 affixes are characterized by the fact that they often trigger changes in the consonant or vowel segments of the base and may affect the assignment of stress [e.g., *-ity, -ive, -ize, -ion*]. (Class 1 affixes often combine with bound roots.) In contrast, Class 2 affixes tend to be phonologically neutral, having no effect on the segmental makeup of the base or on stress assignment [e.g., *-ness, -ful, -ly, -ish*]."

[20] On phonetic grounds alone, it seems impossible to decide whether the consonant following /æ/ in words such as *planet*, *flannel*, *personality*, and *California* is ambisyllabic, or instead falls in the onset of the following syllable. At least in the lexical representation, lax vowels are not allowed in open position in English (or in other Germanic languages such as Dutch and German, for that matter).

Regarding the external (social) independent variable of socioeconomic status (SES), we might expect that the speakers' socioeconomic background would have an effect on the extent of participation in the NCS. Hock & Joseph claim that the Chicago Chain Shift (a subsystem of the NCS) is limited to certain white working-class male groups (Hock & Joseph, 1996:327). In contrast to this generalization, Labov cites Eckert's study in a suburban Detroit high school in which she showed that "the shift was most advanced among females of the upwardly mobile segment of the high school population" (Labov, 1994:189). A more recent study by Eckert & McConnell-Ginet (1995:502-3) found that for later stages in the NCS (i.e., shifts 5 and 6 in Figure 2), it is the groups associated with lower SES who lead in the change.

The idea underlying the selection of the style variants "monologue" (scripted speech) and "dialogue" (spontaneous speech) is that in a dialogue, the amount of attention paid to the speech itself will generally be somewhat less than in monologue speech. In the Labovian definition of style, the relative amount of attention paid to speech plays a central role in the level of formality exhibited in that speech (cf. Labov 1972, Chapt. 3).[21]

The first two linguistic factor groups (grammatical category and membership in *mad, bad, glad* lexical class) were studied to determine whether or not the Columbus dialect shares the lexicalization which has occurred in the Philadelphia dialect. Philadelphia, like Columbus, is classified in the Lower North dialect region (see Wolfram, 1991:85, citing Carver) and is not considered to be participating in the NCS. Thus, Philadelphia offers a reasonable alternate model for Columbus speakers. According to Labov, "[in Philadelphia] all vowels followed by voiced stops are lax, except for those of *mad, bad,* and *glad*, which are always tense" (Labov, 1994:431). To test whether this lexical category operates in the same fashion in Columbus, we coded the tokens for membership in the *mad, bad, glad* class. The variant "other *-ad* word" is included to test for lexical diffusion effects of /æ/ raising on similar words, such as may be taking place with *sad* in Philadelphia (Labov, 1994:431).

Regarding grammatical category, preterite forms of strong verbs ending in nasals, e.g., *ran, swam, began,* show neither tensing nor raising of /æ/ in Philadelphia (Labov, 1994:431; cf. Halle & Mohanan, 1985:107). To determine whether this factor is operational in the Columbus dialect, we coded the data for the following five variants: preterite strong verb, preterite irregular verb.[22] preterite regular (weak) verb, other verbs (non-preterite), and non-verbs.

In the selection of the linguistic variables, we also consider morphological structure to determine whether or not raising and tensing of /æ/ are postlexical processes. Data from Trager (1930)[23] and Halle & Mohanan (1985: 75) suggest that morphological structure is irrelevant for these processes. Halle & Mohanan give the following rule for the realization of /æ/ ("/æ/-Tensing") in "some GA [American English] dialects," as shown in Figure 14. According to Halle & Monahan (pp. 84, 101), this rule operates postlexically, and therefore is not affected by morphological structure.

[21] Strict adherence to Bell's theory of audience design suggests an opposite hypothesis: all radio speech is addressed to the mass target listening audience, resulting in little difference between "monologue" and "dialogue."

[22] Our data yielded no examples of preterite irregular verbs.

[23] Trager's data pertain to Mid-Atlantic states dialects.

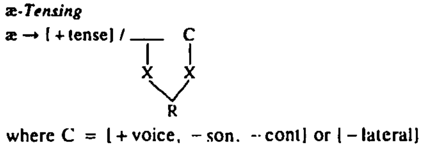


Figure 14. /æ/-Tensing (adapted from Halle & Monahan, 1985:75).

In Kiparsky's (1988) view, however, /æ/ tensing is more complex in several respects. Whereas tensing is restricted to members of lexical categories and applies to the lexical component in Philadelphia and more generally in the Mid-Atlantic area, it is a variable postlexical rule in the Midwest. In Midwestern dialects, the opposition between /æ/ and /a/ "is neutralized in favor of *a* before tautosyllabic *r* by the lexical backing rule" (see Figure 15 below) (402).

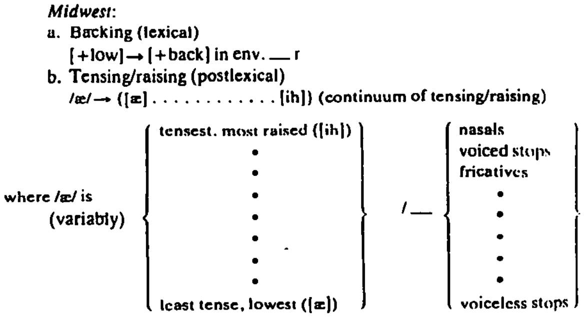

Midwest:
a. Backing (lexical)
[+low] → [+back] in env. ___ r
b. Tensing/raising (postlexical)
/æ/ → ([æ] . . . . [ih]) (continuum of tensing/raising)
where /æ/ is (variably)
tensest, most raised ([ih])
least tense, lowest ([æ])
/ ___
nasals
voiced stops
fricatives
voiceless stops

Figure 15. Lexical and postlexical rules for /æ/ change in Midwestern dialects (adapted from Kiparsky, 1988:402).

In connection with what Kiparsky refers to as a "continuum of tensing/raising," it is worth pointing out that in more extreme cases, the tensing and raising of /æ/ can, e.g., result in a realization of *Ann* which is homophonous with *Ian* (Hock & Joseph, 1996:134). As Kiparsky shows, yet another situation exists in New England, which falls diachronically and phonologically between the situations in the Midwestern and the Mid-Atlantic dialect regions. For the Mid-Atlantic dialects, Kiparsky's analysis is strongly supported by data and analyses presented in Labov 1994 (Chapts. 15-18).

Our selection of the right-hand morphological boundary as an independent variable is meant to test Kiparsky's analysis. Blindness for morphological structure is a feature of most postlexical processes. Therefore, we would predict that if tensing and raising of /æ/ is postlexical, the morphological structure should have no effect on tensing and raising.

Proximity to the right-hand word boundary (measured in terms of syllables) is an additional factor that we considered. The inclusion of this factor group is motivated by the notion that the duration of a given prosodic unit (e.g., syllables, feet) within a word may be related to the length (in syllables) of that word. Perhaps the vowel /æ/ will have a shorter duration the further it is from the end of the word, i.e., the greater the number of syllables that follow it before the word boundary. For example, in the nouns *man*, *mantle*, and *manifold*, all three instantiations of /æ/ occur within the same phonetic environment, yet we might expect that the vowels will have differenct durations: /æ/ in *man* would have the longest duration, while the /æ/ in *manifold* would have the shortest.

An additional factor we considered is stress. Stress has several acoustic correlates, including amplitude, duration, and pitch. The tensing and raising of /æ/ is clearly related to stress; witness the difference in (1) between the realizations of stressed and unstressed /æ/ in the following tokens from our data, where the vowel raises only when it is stressed (primary stress is indicated with an acute accent):

(1) "ánalyze" [e^{ə}] n [ə] lyse
"análysis" [ə] n [Æ] lysis

From this, it seems that the tensing and raising is confined to non-'weak', i.e., non-cliticized words (Labov, 1994:430). As suggested in (1), for polysyllabic words, tensing and raising seem to be restricted to syllables headed by a non-degenerate foot.

Preceding and following phonetic segments have also been shown to affect /æ/ movement. Studies have shown that for both the Mid-Atlantic and Midwestern dialects, there is cross-dialectal variation in the tensing and raising of /æ/ depending on the nature of the following consonant. Labov (1994:430) discusses the considerable differences between the vowel system in New York City and the more restricted Philadelphia system. An example of a phonetically restricted system regarding /æ/ raising can be found in Milwaukee (Chambers, 1995:198-200). According to Chambers, in this dialect the tensing and raising of /æ/ occurs preceding voiced velar segments /g, ŋ/, e.g., *bag* /bæg/ becomes something similar to [beg]. One of the authors, a native of the upper Midwest, notes that this phenomenon is widespread throughout Minnesota and Wisconsin.

Additional studies on the NCS have shown that other preceding and following phonetic segments condition the vowel shift in question. For example, Labov (1994) notes that preceding liquids and following nasals have a significant effect on tensing and raising of /æ/. We will pay particular attention to these environments (preceding liquids and following nasals) in this study, although we coded all preceding and following segments.

Finally, we investigate syllable membership of the consonant following /æ/, because it also has been shown to have an effect on vowel tensing. Labov (1994:432) found an unpredictable distribution of /æ/ tensing in words such as *planet* and *personality*, where it is unclear whether the consonant following /æ/ is ambisyllabic or wholly in the onset of the next syllable (see Footnote 20).

Statistical Methods

For the sociolinguistic aspects of this study, we employ the statistical methods logistic regression and ANOVA, rather than the commonly-used VARBRUL, a statistical technique which is popular among North American sociolinguists. Generally, logistic regression, with which we analyse the tensing (diphthongization/monophthongization) of /æ/, is very similar to VARBRUL. Like VARBRUL, logistic regression has been created for nominal dependent variables, i.e., "discrete choices" (Sankoff, 1987:984). The main advantage of logistic regression is that it makes it easy to analyse the interactions between independent variables; this is very laborious in VARBRUL, and the outcomes are somewhat opaque. Statistics were performed using the SPSS statistical software package for Windows 6.1.3.

FINDINGS: SOCIOLINGUISTIC ANALYSES OF THE TENSING AND RAISING OF /æ/

Analysis of our auditory judgments indicated that /æ/ is variably tensed and raised within the speech of our two speakers. Since we are working within the framework of audience design, we claim that this phenomenon directly reflects the speech of the central Ohio community. That we find /æ/ raising, the initial shift in the NCS, suggests that central Ohio speakers may be participating in the NCS. Alternatively, this /æ/ tensing and raising may be an isolated phenomenon which merely resembles the initial shift in the NCS.

Our auditory judgments of /a/ tokens indicated that this vowel is not undergoing any significant fronting.[24] This dependent variable was included to confirm participation in the NCS, since it is a shift that is unique to the NCS. Lacking evidence for the fronting of /a/, we must conclude that the data gathered thus far are inconclusive for central Ohio's participation in the NCS. They could, however, be evidence of the NCS in its infancy in this area.

Factor Group Effects on the Tensing and Raising of /æ/

We now procede to examine each of the factor group's effects on /æ/ raising and tensing. As was mentioned above, in our analyses of the dependent variable, raising was measured via the height of the vowel (on a five-point scale), and tensing was determined via categorization of the vowel as either diphthongal or monophgthongal (a nominal variable[25]). Table 4 on the following page summarizes the effect of each factor group on raising and tensing (diphthongization): those that are significant are marked "+", and those that are not significant are marked "-." The statistical analyses which yielded these outcomes were analyses of variances (ANOVAs) for raising, and logistic regression for tensing.

Factor groups which were found to be significant for the raising of /æ/ were socioeconomic status (SES), proximity to the right-hand word boundary, stress, preceding liquid segment, following nasal segment, and syllable membership of the following consonant.

Factor groups which were found to be significant for the tensing of /æ/ were SES, preceding liquid segment, following nasal segment, and syllable membership of the following consonant. There was also a significant interaction effect for tensing between SES and style.

The factor "style" (*monologue* versus *dialogue*) does not have a significant effect on the tensing of /æ/, nor on its raising. This negative finding may be due to the relatively low number of observations of dialogue speech (N=39). It may also serve as confirmation of Bell's audience design theory, which predicts little style shifting in radio speech (see Footnote 21).

[24] Instead, /a/ has a small number of slightly-backed variants. This reflects the incomplete merger of the categories /a/ and /ɔ/ (e.g., *cot* and *caught*) in central Ohio and many other parts of the nation. We do not analyze this phenomenon in this study.

[25] A *nominal variable* is a factor which has discrete categories; in this case, our variable "Tenseness" has the categories "diphthong" and "monophthong."

	raising	tensing
EXTERNAL FACTORS		
main effects		
SES (defined by station)	+	+
style (monologue/dialogue)	-	-
interaction effect		
SES and style	-	+
INTERNAL FACTORS		
main effects		
membership in *mad, bad, glad* lexical class	-	-
grammatical category	-	-
right-hand morphological boundary	-	-
proximity to right-hand word boundary	+	-
stress	+	-
preceding liquid segment	+	+
following nasal segment	+	+
syllable membership of the following consonant	+	+

Table 4. The presence ("+") and absence ("-") of significant effects of internal and external factors on the raising and tensing of /æ/ in Columbus.

Interestingly, SES and style exercise a significant interaction effect on the tensing of /æ/, but not on the raising of /æ/. Consider Table 5 below:

	Monologue				Dialogue			
	country		classical		country		classical	
monophthong	71	69.6%	170	84.2%	20	69.0%	9	90.0%
diphthong	31	30.4%	32	15.8%	9	31.0%	1	10.0%
TOTALS	102	100%	202	100%	29	100%	10	100%

Table 5. The interaction effect of style and SES on the tensing of of /æ/. Raw scores and column percentages are shown.

For the country station speaker, the percentage of diphthongs changes minimally between monologue (30.4%) and dialogue speech (31.0%). For the classical station speaker, however, it changes much more (monologue 15.8%; dialogue 10.0%). Remarkably, for the classical station speaker, the proportion of diphthongs decreases going from monologue to dialogue speech, which is opposite the expected effect of greater tensing in casual speech. It should be kept in mind, however, that for the dialogue speech by the classical station speaker, we are generalizing over very few observations (N=10), so the latter effect may well be merely an artefact of the low count.

Neither of the two linguistic factor groups that were used to test Columbus's resemblance to Philadelphia in terms of lexicalization of /æ/ raising and tensing (membership in *mad, bad, glad* lexical class, and grammatical category) has a significant effect on our Columbus data. If these findings are generalizable, then it can be concluded that in Columbus, the tensing and raising of /æ/ has not been lexicalized as it has in Philadelphia.

The latter conclusion is confirmed by the finding that the nature of the nearest following morphological boundary (i.e., /æ/ preceding a Class 1 suffix, /æ/ preceding a Class 2 suffix, /æ/ preceding an inflectional suffix, /æ/ preceding a word boundary) has no meaningful effect on the tensing and raising of /æ/ in our data. The fact that /æ/ raising and tensing appears to be blind to morphological structure suggests that these are postlexical processes.

The number of syllables[26] between /æ/ and the right-hand word boundary has a significant effect on the raising of the vowel. Both qualitatively and quantitatively, raising is inversely proportional to the number of following syllables, i.e., the greater the number of syllables following /æ/, the less likely it is that /æ/ will be raised, and if it is, it will be raised to a lesser degree. Tensing is not sensitive to this variable.

The same overall pattern is found for the next prosodic dimension, word stress. Both qualitatively and quantitatively, raising occurs most in monosyllabic words (e.g., *man*), somewhat less in syllables with primary stress (e.g., *manifold*), and is least likely to occur under secondary stress (e.g., *manifesto*). Example (2) below shows the average height for /æ/ in syllables of each of these stress types, as occurred in our Columbus data. (Primary stress is marked with an accute accent mark, and secondary stress is marked with a grave accent mark.)

(2)

	Example	Avg. height
Stress	word	for stress-type[27]
monosyllabic word	*man*	2.19
primary stress	*mánifòld*	1.83
secondary stress	*mànifésto*	1.55

Although overall stress does not have a significant effect on tensing, there is a significant difference in tensing between monosyllabic words and syllables with secondary stress (/æ/ diphthongizes in 24.9% of monosyllabic words and in only 5.0% of syllables with secondary stress, with B=.7230, SE=.3637, Wald=3.8680, df=1, signif.=.0492, R=.0725, and Exp(B)=2.0606).[28] The fact that the two factor groups "proximity to right-hand word boundary" and "stress" have a significant effect only on raising suggests that the tensing and raising of /æ/ are mutually independent (insofar as the auditory analyses are reliable and the present findings are generalizable). This might be unexpected, due to the

[26] The number of syllables separating the /æ/-containing syllable from the right-hand word boundary is the only *continuous* independent linguistic variable in our set.

[27] "Average height" refers to the average of the scores for height that the /æ/ tokens of this stress-type received, based on the 5-point height scale in Figure 3. A score of "1" indicates no raising; the higher the score, the more raised the token of /æ/ is.

[28] The statistic "B" is the factor weight (the regression coefficient). "Wald" is a statistic with which the significance of the regression coefficient is determined; it has a chi-square distribution. "Exp(B)" is a logistic function, in particular the power to which *e* must be raised to obtain the exponential function.

observations noted above in Table 3, which showed that height and tenseness, as dependent variables, do interact.

We studied both the left-hand and right-hand segmental environments. Both turn out to have significant effects on the raising and the tensing of /æ/. Regarding the left-hand environment, a liquid causes significantly less raising as well as less tensing than any other (natural class of) consonant. Conversely, a following nasal triggers significantly more raising and tensing than any other consonant. As Hock & Joseph (1996:134) summarize, the Northern Cities realization of /æ/ tends to be strongly nasalized (at least when followed by a nasal segment). In this respect, the Columbus dialect again resembles those spoken in cities participating in the NCS. Preceding /l/ is a relatively disfavoring environment for /æ/ raising and tensing in the NCS (Labov, 1994:458). We found this environment to have the same effect in our data. Interestingly, the dimensions "liquid versus other consonant preceding" and "nasal versus oral consonant following" are among the independent linguistic variables that turned out to have strong effects on the tensing and raising of /æ/ in Philadelphia, which is not in the NCS (Labov, 1994:512).

Although we do not further investigate the effects of left-hand and right-hand segmental environments, our data would permit testing for such potentially relevant factor groups. For example, we are in a position to test the claims regarding the conditioning right-hand environment for the tensing of /æ/ in Halle & Mohanan's analysis (reproduced in Figure 14 above). However, even a superficial inspection of the relevant crosstabulations (not presented here) makes it clear that in our data there are no (classes of) consonants before which the tensing and raising of /æ/ is blocked.

The last linguistic dimension studied in connection with the tensing and raising of /æ/, the syllable membership of the following consonant, turns out to have a significant effect on both tensing and raising. In connection with the syllable membership of the following consonant, we distinguished four conditions as follows. The following consonant is:

- either ambisyllabic or part of the onset of the next syllable;
- one tautosyllabic consonant;
- two tautosyllabic consonants; or
- three tautosyllabic consonants

The relative importance of these four conditions is not entirely identical for raising versus tensing. For raising, the ordering of importance is:

	3 tautosyll. C >	1 tautosyll. C >	2 tautosyll. C >	ambisyll./next syll.
average height:	3.24	2.02	1.92	1.66

while for tensing, the ordering of importance is:

	3 tautosyll. C >	2 tautosyll. C >	1 tautosyll. C >	ambisyll./next syll.
percent diphth.:	66.7 %	23.3 %	19.3 %	11.0 %

Although both the strongest and the weakest effects in both dimensions are related to the same factors, for this factor group (which is related to syllable weight) the raising and the tensing of /æ/ are not entirely identical in their conditioning.

Relative Weight of Factor Group Effects for the Tensing and Raising of /æ/

Finally, we examine the relative importance of the eight independent linguistic variables to the raising and tensing of /æ/. We investigate which variables play the most meaningful role on the overall extent of raising and tensing, as well as their weights.

To establish the most important factors for raising, we ran a multiple regression analysis.[29] The main outcomes are presented in Table 6.[30]

variable	B	SE B	Beta	t	signif.	MR	R^2
-/+nasal foll.	1.2224	.0751	.6475	16.282	.0000	.6836	.4673
stress	.2462	.0592	.1655	4.161	.0000		
constant	-.2545	.1726		-1.475	.1413		
F=149.1551	df=2,340	signif.=.0000					

Table 6. Findings from multiple regression analysis regarding the internal factors most important for raising. Factor groups entered into the analysis included the eight internal factors listed both in the section "Selection of Independent Variables" and Table 4.

The nasal/oral nature of the following consonant turns out to be the main predictor for the raising of /æ/. The nature of this effect is, as discussed above, that significantly more raising of /æ/ occurs preceding a nasal. The second and only other significant predictor is word stress. The relationship between stress and raising in this multiple regression equation is positive: stressed monosyllables induce the most raising, syllables with primary stress induce less raising, and syllables with secondary stress induce the least raising (see Example (2)).

These two significant factor groups together accont for 47% (R^2) of the variance in the height of /æ/ in the data for these two speakers.

To establish the most important factors for tensing, we applied logistic regression[31] to our data, with all eight internal factors groups as predictors (in view of the nominal nature of this variable, logistic regression is most appropriate method). The most important outcomes can be found in Table 7.

[29] With forward inclusion. At each step, the criterion for a variable to be entered is for its F-to-enter to have a probability smaller than .05, while a variable with a probability greater than .10 is removed.

[30] "Beta" is the weight of the independent variable that goes into the equation for the standard scores (the "z-scores") of the dependent variable. (In the equation with z-scores, the intercept , or the constant, is zero.) "F" indicates the overall goodness of fit for this analysis.

[31] With forward inclusion. At each step, the criterion for a variable to be entered is for its score statistic to have a probability smaller than .05, while a variable whose likelihood ratio has a probability greater than or equal to .10 is removed.

	chi-square	df	sign
-2 log likelihood	317.367		
model chi-square	37.761	4	.0000
goodness of fit	341.142		

variable	B	SE	Wald	df	sign	R	Exp(B)
-/+nasal follows	.5138	.1488	11.9244	1	.0006	.1672	1.6716
syll.membership foll. C			11.2008	3	.0107	.1210	
1 vs.3 tautosyll.C's	.3028	.2232	1.8413	1	.1748	.0000	1.3537
2 vs.3 tautosyll.C's	.1516	.2649	.3275	1	.5672	.0000	1.1637
ambisyll./next syll.vs 3 tautosyll C's	.7955	.3289	5.8494	1	.0156	.1041	2.2156
constant	1.0169	.1740	34.1706	1	.0000		

Table 7. Findings from logistic regression analysis regarding the internal factors most important for tensing. Factor groups entered into the analysis included the eight internal factors listed both in the section "Selection of Independent Variables" and Table 4.

In the displays above, "B" is indicative of the factor weight. For the factor group "syllable membership of the following consonant," the contrast between a situation in which the following consonant is ambisyllabic versus one in which /æ/ is followed by three tautosyllabic consonants does have a highly significant effect on the diphthongization of /æ/. For that reason the factor group does as well. This is true in spite of the fact that the two contrasts within this factor group which have the smallest weights (namely one versus three tautosyllabic consonants, and two versus three tautosyllabic consonants) do not play a significant role in the diphthongization of /æ/.

The overall outcome of this analysis is that only two out of the eight internal factors have a meaningful effect on the diphthongization of /æ/: following nasal segment and syllable membership of the following consonant. Both are properties of the following consonant, the first one regarding the oral versus nasal nature of the segment and the second one regarding syllable membership.

A striking aspect of the findings for the multiple regression (height, i.e., raising) and the logistic regression (diphthongization, i.e., tensing) is the fact that the oral/nasal nature of the following consonant emerges as the main predictor for both dimensions in the process of the tensing and raising of /æ/. A better understanding of the process of coarticulation of vowels with following nasals might be instrumental in explaining this finding.

The only other significant predictor emerging from the analysis of raising is stress. Regarding tensing, the only other significant predictor selected is syllable membership of the following consonant. These two factors are not identical, but they are not entirely unrelated either, as both stress and relative syllable membership pertain to prosodic organization.

QUESTIONS FOR FUTURE RESEARCH

This pilot study leaves many interesting questions for future research, most of which can be addressed through further examination of our current corpus, requiring no further data collection.

In terms of internal factor groups, the effects of the preceding and following phonetic segments can be explored in much greater detail. For example, for some central Ohio speakers, a following palatal segment will have the effect of tensing (and, in some cases, raising) an otherwise lax vowel, e.g., *crash* becomes [kɹeʃ], *treasure* becomes [tɹeʒɚ], *fish* becomes [fiʃ], and *bush* becomes [buʃ]. We might explore the extent to which preceding or following palatals have a favoring effect on the the tensing and raising of /æ/. More generally, what is the effect of place of articulation on tensing and raising? Is it comparable to Labov's finding that following palatal and apical segments favor raising, whereas following labial and velar segments disfavor raising in the NCS (1994:458-9)? Does the voice specification of the preceding or following consonant have a systematic effect on tensing and raising? Does a preceding single consonant versus a preceding consonant cluster have an effect? And what about the sonority value of the following consonant?

Another question to be addressed in further research is whether or not the linguistic factor groups we studied are mutually independent. In other words, do their *interactions* exert any significant effects on the raising and the tensing of /æ/? In particular, we would be interested in the effect of the following interactions:

- between the factor group "number of syllables to the right-hand word boundary" and the factor group "stress"
- between the factor group "number of syllables to the right-hand word boundary" and the factor group "syllable membership of the following consonant"
- between the factor group "stress" and the factor group "syllable membership of the following consonant"
- between the factor group "liquid versus other consonants preceding" and the factor group "nasal versus other consonants following"

The interaction effects can be studied for raising (using ANOVA) as well as for tensing (using logistic regression).

For both the analyses of the main effects and for the analyses of the interaction effects, it would be preferable to treat the linguistic conditions underlying each internal factor group as "repeated measure(ment)s." This refinement will be implemented in the next stage of our research.

If we are to observe the spread of the NCS or any other sound change through the lexicon of central Ohio speakers, then lexical frequency effects could also be examined. While we may expect frequency effects to have the most impact on deletion or reduction processes, we do not want to rule out their effect on vowel tensing without further study. Frequently-used lexical items might tend to incorporate vowel changes first. On the other hand, unshifted vowels in these same high-frequency lexical items may serve as important sociolinguistic markers of group identity, and therefore be more resistant to change. In studying this factor, we would need to make a clear differentiation between lexical frequency defined generally and lexical frequency as an artefact of our data. For example, the phrase *Club Dance* has high frequency in "Red's" promotional banter, and the word *Mastercard* appears frequently throughout "Daniel's" fund-raising appeals. Neither of these phrases occurs frequently in the general lexicon.

Other studies clearly indicate that socioeconomic status can interact with gender (e.g., Eckert, 1986; Eckert & McConnell-Ginet, 1995; Hock & Joseph, 1996:327). In a future study, we hope to include speech from female speakers.[32]

[32] The choice of male radio announcers in this pilot study was driven by the fact that we were unable to identify any female announcers who were natives of central Ohio.

Refinements in our instrumental analysis of vowel tensing and raising are also in order. A more accurate characterization of the nature of diphthongs will require a more precise methodology for measuring formant movement throughout the vowel. For example, we must consider the placement and slope of formant changes within vowels.

Since our auditory and acoustic analyses are based on radio-broadcast speech, we might wonder whether this form of sound transmission has any systematic distortion effect on the acoustic signal. In other words, how suitable is radio speech for acoustic analysis? We could attempt to assess whether or not the acoustics of natural speech vary significantly from radio broadcast speech by comparing laboratory recordings of both speakers' speech to our radio tape-recordings.

On the basis of some of the aspects of the present study, a larger scale sociolinguistic study of possible shifts in the Columbus vowel system could be designed, based partly or not at all on radio speech. This follow-up study could include analysis of the other vowels which make up the NCS. In particular, /a/ merits further attention. Our current study revealed little fronting of /a/, but perhaps any subtle movement occurring could be better detected by more careful instrumental measures. Any study of the shift of /a/ must also take into account the collapse of /a/ and /ɔ/, a merger which divides Ohio from southwest to northeast, roughly along Interstate 71, which runs directly through Columbus. In this merger, the vowels in pairs such as *Don* and *dawn,* and *cot* and *caught* are pronounced the same by speakers southeast of this dividing line (Labov, 1996). If the /a/-fronting from the NCS takes hold in central Ohio (an area where the merger has uncertain status) , to what extent will it impact words such as *cot* and *caught* in the same way?

CONCLUSIONS

This pilot study yields the following provisory conclusions.

In the central Ohio dialect, the vowel /æ/ shows variation in tensing and raising simliar to the Northern Cities Shift. However, the vowel /a/ is not fronted as it is in the NCS. Thus, we cannot state conclusively that the variation of /æ/ observed in our data indicates participation in the NCS.

The study of the external conditioning factors (SES and style) show that the tensing and raising of /æ/ is socioeconomically stratified, occuring to a greater extent in the speech of speakers of relatively lower SES. However, the available data do not to allow any conclusions as to the stylistic conditioning of the process.

The study of the internal conditioning factors showed that tensing of /æ/ and raising of /æ/ are related but not mutually dependent. The nasal/oral nature of the following consonant is the factor with the strongest influence on both the raising and the tensing of /æ/. Secondary factors are prosodic in nature: stress (for raising) and syllable membership of following consonants (for tensing). The tensing and raising of /ae/ probably apply postlexically.

As these results indicate, there is considerable externally- and internally-influenced variation in the production of /æ/ for Columbus speakers. Further study of the vowel systems of central Ohio speakers will shed additional light on the status of their dialects, and the relationship of the variation in their vowel systems to the vowel shifts taking place in neighboring dialects.

REFERENCES

Bell, Allan. (1984) Language style as audience design. *Language in Society*, June, 13:2, 145-204.

Chambers, J.K. (1995) *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Cambridge, MA: Blackwell Publishers.

Eckert, Penelope. (1986) The roles of high school social structure in phonological change. CLS presentation, Chicago.

Eckert, Penelope & Sally McConnell-Ginet. (1995) Constructing meaning, constructing selves. In Kira Hall & Mary Bucholtz (eds.), *Gender Articulated: Language and the Socially Constructed Self.* New York: Routledge, 469-507.

Fujimura, Osamu & Kakita, Y. (1975) Remarks on Quantitative Description of the Lingual Articulation. In B. Lindblom & S. Öhman (eds.), *Frontiers of Speech Communication Research.* New York: Academic Press, 17-24.

Gordon, Matthew. (1996) Urban Sound Change Beyond the Cities: The Spread of the Northern Cities Chain Shift. NWAVE 25 presentation. Las Vegas, Nevada, October.

Halle, Morris & K. Mohanan. (1985) Segmental phonology of Modern English. *Linguistic Inquiry*, 16:1, 57-116.

Hock, Hans Henrich & Brian D. Joseph. (1996) *Language History, Language Change, & Language Relationship.* Berlin: Mouton de Gruyter.

Ito, R. (1996) The Northern Cities Shift in Rural Michigan. NWAVE 25 presentation. Las Vegas, Nevada, October.

Kiparsky, Paul. (1988) Phonological change. In F. Newmeyer (ed.), *Linguistics. The Cambridge Survey.* Volume 1. Cambridge: Cambridge University Press, 363-415.

Labov, William. (1972) *Sociolinguistic Patterns.* Philadelphia: University of Philadelphia Press.

Labov, William. (1991) The three dialects of English. In Penelope Eckert (ed.), *New Ways of Analyzing Sound Change.* New York: Academic Press, 1-44.

Labov, William. (1994) *Principles of Linguistic Change.* Cambridge, MA: Blackwell Publishers.

Labov, William. (1996) The organization of dialect diversity in North America. ICSLP4 presentation. Philadelphia, October.

O'Grady, William, Michael Dobrovolsky, & Mark Aronoff (eds.). (1997) *Contemporary Linguistics: An Introduction.* 3rd edition. New York: St. Martin's Press.

Sankoff, David. (1987) Variable rules. In U. Ammon, N. Dittmar & K. Mattheier (eds), *Sociolinguistics / Soziolinguistik. An international handbook of the science of language and society / Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft.* 2nd Volume. Berlin: Mouton de Gruyter, 984-97.

Trager, G. (1930) The pronunciation of 'Short A' in American English. *American Speech*, 5, 396-400.

Van de Velde, Hans & Roeland van Hout. (1996) Radio Broadcasts as a Source for the Study of Language Change. NWAVE 25 presentation. Las Vegas, Nevada, October.

Van de Velde, Hans, Roeland van Hout & Marinel Gerritsen (1996) Watching Dutch change. A real time study of variation and change in standard Dutch pronunciation. Manuscript, University of Nijmegen.

Wolfram, Walt. (1991) *Dialects and American English.* New York: Prentice-Hall.

Zeller, C. (1993) The investigation of a sound change in progress: /æ/ to /e/ in Midwestern American English. NWAVE 22 presentation. University of Ottawa, October.

Syntactically-Governed Accentuation in Balinese*

Rebecca Herman
rherman@ling.ohio-state.edu

Abstract: In Balinese there is a consistency of alignment between F0 peaks and particular syntactic positions such as "final syllable of the head of the phrase" or "final syllable of the phrase." This becomes apparent from F0 measurements taken from sentences recorded from a Balinese speaker which include measurements from sentences with different syntactic constructions and different length words in each syntactic position. Thus, the placement of F0 peaks in Balinese is not distinctive and in fact, there is no word-level accentuation in Balinese. Rather, placement of F0 peaks occurs at the phrasal level and hence serves a delimitative function.

INTRODUCTION

Field-work on Balinese suggests that there is no word-level stress or accentuation in this language at all, but that accentuation does exist at the phrasal level and is governed by syntactic principles. Thus, Balinese differs from tone-languages such as Cantonese (with lexically contrastive use of fundamental frequency (F0)), from classic pitch-accent languages such as Serbo-Croatian and Japanese (with lexically determined placement of accent), and from classic stress-accent languages such as English and Italian (with pitch accents chosen from an inventory of intonational morphemes and associated to lexically contrastive locations in words). Instead, the accentuation in Balinese resembles languages like Korean (as described by Jun, 1993) and French (as described by Jun and Fougeron, 1995), in which the "pitch-accent," or localized F0 excursion, serves a delimitative purpose.

This paper describes the evidence for such a characterization of accentuation in Balinese. The argumentation is as follows. First, it will be argued that there is no word-level stress or accentuation in Balinese. Then, the syntactic principles governing accentuation will be illustrated by showing a consistency of alignment of F0 peaks with syllables in particular syntactic positions. The discussion of these syntactic principles will be divided into discussion of accentuation in single-word subjects, in more complex subjects, in predicates, and in clauses. This will be followed by a discussion of sentence-level accentuation, including the use of F0 in question-formation and in focus.

METHODS

The corpus of data examined here was elicited and recorded during field-work with a Balinese speaker. The sentences were recorded in a quiet room in the speaker's home using a Sony "professional" portable tape recorder and digitized using Waves™. When F0 measurements are reported, they are measured at the highest point in the vowel for syllables with a clearly visible F0 peak and at the center of the vowel for vowels without a clearly visible F0 peak, as shown in figure 1. The only acoustic correlate studied here was F0 variation, since there were clear F0 events in the recorded speech. This is not to imply that

* Funding for this project was provided by an NSF Graduate Student Fellowship Research Allowance. I wish to thank Ayu Krisna Dewi Wijaya for being such a great consultant and for making field-work so much fun, and thanks to her and to I Ketut Putra Erawan for welcoming me into their home. Thanks also to Mary Beckman for providing suggestions about elicitation methods and for comments on the paper.

other potential correlates, such as duration or intensity patterns, are not important components of accentuation in Balinese. Rather, the corpus does not include materials appropriate to examine these phonetic cues in production. Perception tests manipulating these factors would be needed to determine the perceptual importance of each factor, but for now, in this production study, the factor under investigation is F0.

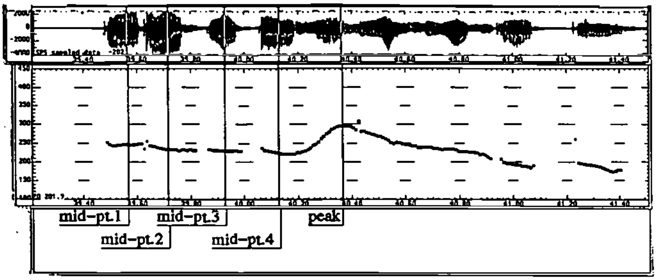


Figure 1. A sample F0 trace and speech waveform, indicating where F0 was measured for the sentence [pɨrabotane luwɯŋ luwɯŋ gati] :"his furniture is very nice." The vertical lines marked "mid-pt. 1," "mid-pt.2," "mid-pt.3," and "mid-pt.4" are at the mid-points of the vowels in the syllables [pɨ], [ra], [bo], and [ta] respectively, and the vertical line marked "peak" is marking the highest point in the F0 peak on the syllable [ne].

The speaker is a 27 year old woman from Tabanan, in Southern Bali. Balinese is her primary language, learned at home. Indonesian was her language of education, and she also speaks Javanese and English. In several of the Indonesian languages, including Balinese, there are ways of expressing relative social status between speakers (as determined by caste, skills, age, and wealth) through the use of language (Stevens, 1965; Barber, 1977; Ward, 1973; and Geertz, 1972). Respect for the addressee is shown by the speaker by the use of High language and lack of respect for the addressee is shown by the speaker by the use of Low language. The levels of language are expressed mainly by the choice of lexical items and not by phonology, morphology, or syntax. Although most of the vocabulary is neutral, it has been estimated that several hundred words, including many of the most commonly used words, have forms in more than one status set (Ward, 1973; Barber, 1977). Changes in modern Balinese society, including a breakdown of the traditional caste system, have raised problems for the use of levels of politeness in language. It has been observed that the traditional norms of language usage with respect to levels of politeness are on the decline (Shadeg, 1977), and this observation is confirmed by the impressions of the speaker in this study. The speaker used in this study controls both the High variety and the Low variety of Balinese. For consistency's sake, and since the situation in field-work is an unnatural discourse situation in that the addressee is the field-worker (and not a speaker of Balinese), for this study the speaker was asked to speak as she would to friends. This request resulted in a blend of High and Low vocabulary, or a type of "Mid" speech.

Segmental effects were not controlled for, but on the whole did not affect the outcome of the F0 measurements, as can be seen by the tight clustering of F0 values in the graphs throughout the paper, regardless of segmental content. For reference, the vowel inventory of Balinese consists (phonemically) of /u i o e a ɨ/. The high tense vowels /i u/ have lax counterparts [ɪ ʊ] which occur in word-final closed syllables, which is shown both by distribution and by alternations. The mid tense vowels /e o/ have the lax counterparts [ɛ ɔ] which occur in word-final closed syllables and also preceding word-final closed syllables with [ɛ] and [ɔ], again shown by both distribution and alternations. The vowels are transcribed phonetically in this paper, not phonemically. There are also restrictions on which vowels may co-occur within morphemes. The high vowels may co-occur with each other in a morpheme (both tense and lax counterparts) and the mid vowels may co-occur with each other in a morpheme (both tense and lax counterparts), but the high vowels may not occur in the same morpheme with the mid vowels. On the other hand, /ɑ/ and /ɨ/ may co-occur in a morpheme with either high vowels, mid vowels, or themselves. Syllables in Balinese are of the form V, CV, or CVC. The consonant inventory of Balinese is shown in figure 2 for reference.

	labial	coronal	dorsal	glottal
stop	p b	t d	k g	ʔ
fricative		s		h
affricate		ʧ ʤ		
nasal	m	n	ŋ	
approximant	w	y r l		

Figure 2. The Consonant Inventory of Balinese

THE LACK OF WORD-LEVEL ACCENTUATION

It is theoretically impossible to prove that some entity does not exist. Therefore, it is impossible to prove that word-level accentuation does not exist in Balinese. However, if word-level accentuation in some form did exist, one might expect to find certain indications of it. First of all, there might be minimal pairs of words in the langauge which contrasted only in accentuation pattern. There are no such minimal pairs in Balinese.

Furthermore, the native speaker of the language would be expected to have some intuitions about the prominence of particular syllables relative to the other syllables in a lexical item. Thus, anecdotal evidence suggests that native speakers of English, even as school-children, can tap out stress patterns of English words, indicating with stronger taps which syllable is stressed. Anecdotal evidence suggests that this is an impossible task in Balinese. The speaker was asked to clap out words with her hands, giving a stronger clap for more prominent syllables (with a demonstration from English). However, this attempt at uncovering word-level stress or prominence in Balinese was unsuccessful, because not only did the speaker not clap her hands more strongly on any particular syllable, but also she could not quite see the point of the exercise.

If a language had word-level accentuation, it might also be expected that words in isolation would display an accentuation pattern, and would show analogous accentuation to that which is present in sentences. Thus, further evidence against word-level accentuation in Balinese comes from a comparison of recordings of words in isolation with recordings of words in sentences. In words said in sentences, there are clear F0 peaks aligned with certain syllables (the principles of which will be discussed below). In words in isolation, on the other hand, it is possible to have a word with a completely flat F0 pattern (allowing for micro-segmental perturbations), although it is actually difficult to record a word in isolation without eliciting what appears to be a list intonation, with a sharp rise in F0 on the final syllable of each word (except the last). Even attempts to record single words in isolation often show this sharp rise on the last syllable, which seems to be analogous to what is known as "continuation rise" in other languages, as the speaker seems to be indicating that she is willing to continue to the next word, or that further material follows. Some examples of "accent-less" words recorded in isolation are given in figure 3, although such words are atypical examples of utterances and most of the words recorded during the same session do show a continuation rise. Some examples comparing words said in isolation to the same word said in a phrasal context are given in figure 4, although again the accent-less words would be the atypical case.

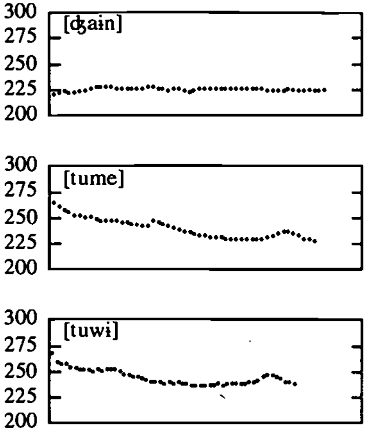


Figure 3. F0 traces of words in isolation, showing no F0 accent (which is atypical and difficult to elicit).

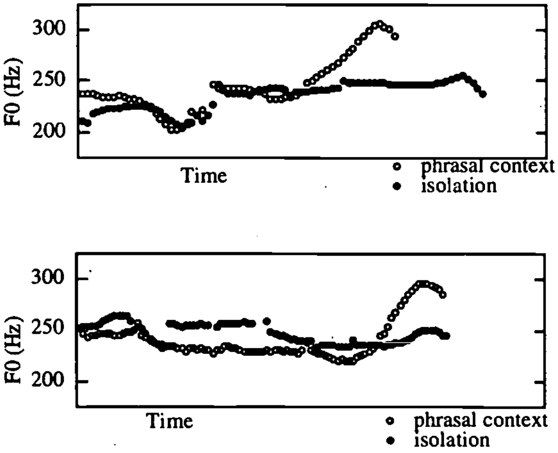


figure 4. F0 traces of the same word recorded in isolation overlaid on F0 traces of words extracted from a phrasal context. The top graph is the word [baʤune] "the clothes" and the bottom graph is the word [pirabotane] "the furniture."

Furthermore, the brief descriptions of word-level accentuation or stress in reference grammars may be taken as supporting evidence. The problem with authors' descriptive accounts of sress or accent is the interference from the author's native language and the consequent reluctance of speakers of stress-accent languages to posit a stress-less lexical system. For example, Barber (1977, p.15) does say that "There is no strong word-stress in Balinese in ordinary speech, there is only a slight variation in loudness and energy between the syllables of a sentence." This will be taken as supporting evidence for the hypothesis that there is no word-level accentuation in Balinese, despite the fact that, as a speaker of a stress-accent language, presumably there is some reluctance on Barber's part to maintain the assertion completely, since he goes on to describe word-stress as "In words of more than two syllables (not counting suffixes), the penultimate syllable is stressed unless the vowel is e." [which is transcribed as [ɨ] here] (Barber, 1977, p. 15)

Further supporting evidence comes from Indonesian and Javanese, languages which are closely related to Balinese. There have been several conflicting accounts of word-stress in Indonesian, leading Odé (1994) to rethink the notion "word-stress" in Indonesian. She provides a summary of the literature on Indonesian word-stress, in which she says that while previous authors all agree that word stress is not distinctive, they disagree on whether word-stress is fixed on the penultimate syllable, on whether word stress is fixed on the final syllable, on whether a schwa is stressable, and on whether pitch is a cue for word-stress. She shows that the literature on word stress in Indonesian is not based on perceptual evidence, and that according to her study,

> ...prominence in Indonesian cannot be described in terms of stressed or accented syllables as described in the literature. Therefore, the syllable does not seem to be the level on which prominence must be studied. (Odé, 1994, p. 63)

Her study focusses instead on prominence on the prosodic phrase level. Odé's study will be taken as supporting evidence for the claim that there is no word-level stress in Balinese, and that accentuation in Balinese must be studied at the phrasal level. Similarly, there is supporting evidence from Javanese, which is also related to Balinese. Horne (1961), in a textbook on Javanese, gives a description of its accentuation patterns which makes it clear that there is no word-level accentuation, although there are phrase- or sentence-level patterns of accentuation. She writes,

> Javanese, unlike English, lacks word accent. It makes no difference which syllable of a Javanese word gets the loudest stress. Sentences in Javanese, on the other hand, have certain characteristic accent patterns. (Horne, 1961, p. xxvi)

Although of course speculation about the lack of word-level stress in related languages is not proof that there is no word-level stress in the language in question, it will nonetheless be used as another piece of supporting evidence, even though it would make a weak argument on its own.

Thus, although it is in fact impossible to prove the non-existence of some entity, the pieces of evidence described above– lack of minimal pairs, lack of speaker intuitions, flat F0 patterns on words in isolation, previous accounts of stress (or lack thereof), and studies of stress in very closely related languages– are all suggestive of a lack of word-level prominence in Balinese.

ELICITATION APPROACH

Odé (1994, p. 63) concludes on the basis of perception experiments that "prominence in Indonesian cannot be investigated in the way we are accustomed to in intensively investigated languages with word stress such as, for instance, English..." This is the attitude that will be adopted here. Thus, instead of comparing lexical items to each other in order to determine the relative prominence of various syllables with respect to each other, as might be useful in a stress-accent language, the items under comparison in this study will be sentences with various syntactic structures and sentences containing varying numbers of syllables in each syntactic position. For example, sentences with one-word nominal subjects can be compared to sentences with more complex nominal subjects to see if there is an F0 peak on the same syllable in the head noun regardless of the complexity of the phrase. Or, sentences with monosyllabic nominal subjects can be compared to sentences with di- or tri-syllabic nominal subjects to see if the F0 peak aligns with a particular syllable in the subject, and how that syllable is determined. These types of comparisons in the present corpus show that there is an F0 rise on the "same" syllable in the sentence, syntactically speaking, regardless of other factors such as the length of individual words or the syntactic category of individual words. This consistency of alignment of F0 peaks with particular syntactic positions is shown by taking the mean F0 values from a particular syllable (such as, say, the final syllable of the head of a phrase) and comparing that mean to the mean F0 values from other syllables (such as, say, all pre-final syllables of the head of a phrase). If the means are significantly different from each other, then it will be argued that there is an F0 peak aligned with a particular position in the sentence. It will be shown that the relevant positions or categories to which F0 peaks align are not things like "noun" or "verb" but rather things like "head of the phrase" or "final syllable in the phrase." Hence, the claim can be made that accentuation in Balinese is syntactically governed.

ACCENTUATION ON SINGLE-WORD SUBJECTS

To begin, sentences with single-word nominal subjects can be elicited. The subject nouns can contain any number of syllables, from two up to about six. Examples include the two-syllable word [yehe] "the water," the three-syllable word [paʤiŋe] "the umbrella," the four-syllable word [sɨpedane] "the bicycle," the five-syllable word [pɨrabotane] "his furniture," and the six-syllable word [matematikane] "the math." No matter how many syllables the word contains, though, there is always an F0 rise on its final syllable. One such token is shown in figure 5. The peak is on the third syllable of the subject noun. (The rise on the last syllable of the sentence seems to be the continuation rise discussed above.)

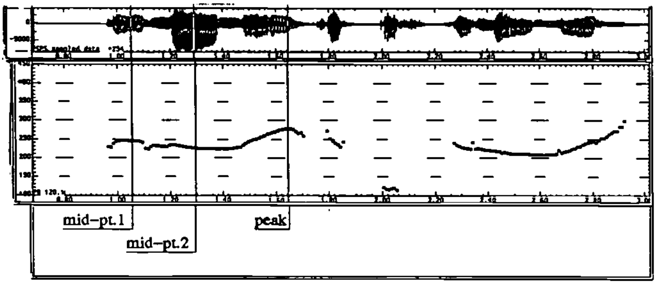


Figure 5. A sample F0 trace and waveform of a one-word nominal subject with a peak on the final syllable of the noun. The sentence shown here is [limane ʤɨpɨs ʤɨlanan] "his hand was caught in the door." The vertical lines marked "mid-pt. 1" and "mid-pt. 2" are marking the mid-points of the vowels in the syllables [li] and [ma] respectively, and "peak" is marking the highest point of the peak in the third syllable [ne].

Figure 6 shows the means (with standard error bars) of all of the pre-final syllables for all of the tokens (measured at the center of the vowel) compared to the mean of all of the final syllables (measured at the highest point of the peak). An unpaired t-test shows that the F0 values of pre-final and final syllables are statistically different at the .05 level. The point on this graph averages over 4 examples with 2-syllable subjects, 3 examples with 3-syllable subjects, 3 examples with 4-syllable subjects, 2 examples with 5-syllable subjects, and 1 example with a 6-syllable subject. Thus, since the means of pre-final syllables are well-separated from the means of final syllables and are in fact statistically different, and since there is a tightly compacted, non-overlapping error, it can be seen that there is an F0 peak aligned with the last syllable of nominal subjects, no matter how many syllables those nouns contain.

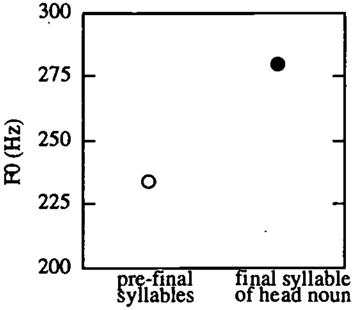


Figure 6. Pre-final vs. final syllables in 1-word nominal subjects. The sentences used are listed in Appendix 1a.

The same phenomenon is observed in each conjunct of conjoined noun phrases. For example, two phrases can be conjoined with [aʤaʔ] "and" or with [napi] "or." In these cases, there is an F0 peak aligned with the last syllable of the nominal subject in each conjoined phrase.

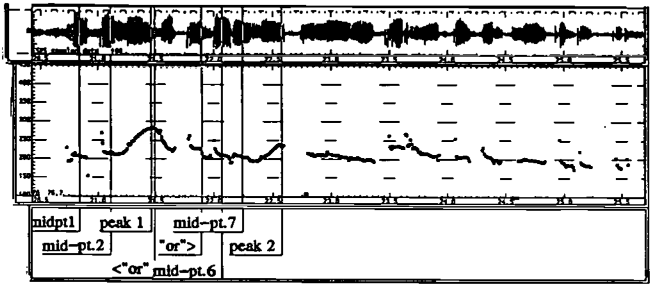


Figure 7. A sample F0 trace and waveform of a sentence with conjoined nominal subjects. The sentence shown here is [bapane napi bɨline ane lakar natıhın iyɨ kɨ dɔktir] "It is his father or his brother that will drop him off to the doctor." "mid-pt. 1" is for [ba] , "mid-pt. 2" for [pa], "mid-pt. 6" for [bɨ] and "mid-pt. 7" for [li] while "peak 1" and "peak 2" mark the highest points in the peaks on the last syllable [ne] of each conjunct. The angled brackets marked "or" show the edges of the word [napi] meaning "or."

The F0 peak on the final syllable of each conjunct is higher than the F0 values of all of the pre-final syllables in that conjunct, although the F0 peak on the final syllable of the second conjunct is lower than the F0 peak on the final syllable of the first conjunct. The fact that the second peak is lower than the first may be indicative of sentence-level declination. The points in figure 8 average over two examples with a 3-syllable first conjunct and a 3-syllable second conjunct and one example with a 2-syllable first conjunct and a 3-syllable second conjunct. Unpaired t-tests show that the last syllable is statistically different than the pre-final syllables at the .05 level for the first conjunct and for the second conjunct as well. Thus, there is a consistency of alignment between the F0 peak and the final syllable of the subject noun, as shown by the well-separated means of pre-final and final syllables of each conjunct, regardless of the length of the subject noun in each conjunct.

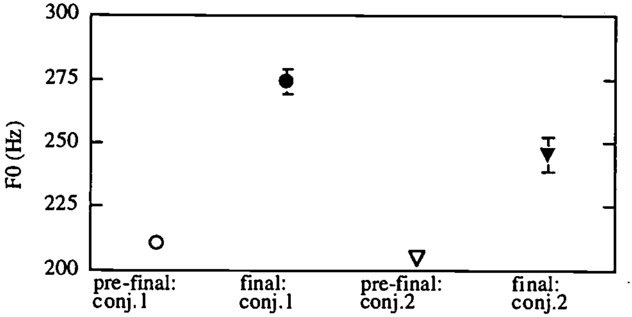


Figure 8. Pre-final vs. final syllables in conjoined single-word nominal subjects. The sentences used here are shown in Appendix 1b.

ACCENTUATION IN MORE COMPLEX SUBJECTS

In more complex subject phrases longer than one word, the same pattern is observed in the head of the phrase, but there is an additional F0 peak later in the phrase. In these cases, then, not only is there an F0 peak on the last syllable of the head of the phrase (as described above for one-word subjects) but there is also an F0 peak on the last syllable of the phrase itself– again, regardless of the number of syllables involved. Thus, there may be any number of syllables in the head of the phrase, but the first F0 peak in the subject will always be on the last syllable of the head. Similarly, there may be any number of intervening syllables between the head of the phrase and the end of the phrase, but the next F0 peak in the subject phrase will occur on the last syllable of the phrase itself. Figure 9 shows a token with a noun phrase (NP) containing 2 adjectives. The first F0 peak is on the third syllable, which is the final syllable of the head noun. The next F0 peak is on the seventh syllable, which is the final syllable of the NP (that is, the final syllable of the second adjective).

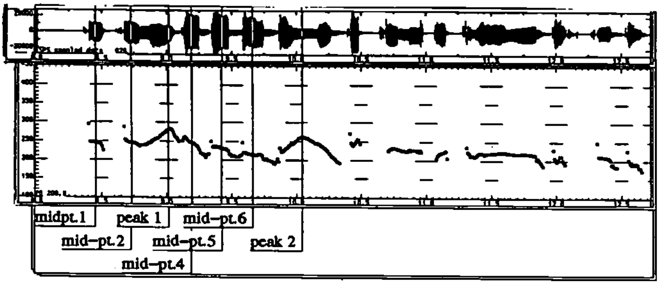


figure 9. A sample F0 trace and waveform of a more complex NP. This is a sentence with two adjectives following the head noun in the subject. The sentence shown here is [ʧiʧiŋɨ gɨde badiŋ nyɨgʊt anaʔe tuwɨ ŋidɪh idɪh] "the big black dog bit the old beggar person." The lines marked "peak 1" and "peak 2" mark the highest points of the peaks on the syllables [ŋɨ] and [diŋ], which are the final syllables of the noun [ʧiʧiŋɨ] and the adjective [badiŋ].

Examples of constructions examined include genitive constructions, adjective phrases, and pre-posed stative verbs. Genitives are formed with the suffix [-n]. Such sentences are of the form [[noun-of] mine] or [noun-of [[noun-of] mine/his]]]. For example, the phrase [bapan timpal tiyaŋe] means "the father of the friend of mine." This example would have the two-syllable head noun [bapan] "the father of" and four syllables (namely, [tim.pal. ti.ya.]) intervening between the final syllable of the head noun and the final syllable of the phrase (which would be the [ŋe] of [tiyaŋe] in this case). The points in figure 10 average over 10 examples of the form [[noun-of] mine], seven of which had 2-syllable nouns, two of which had 3-syllable nouns, and one of which had a 4-syllable noun. The points in figure 10 average over four examples of the form [noun-of [noun-of mine/his]], all of which had 2-syllable head nouns. There was a peak on the final syllable of the head noun and on the final syllable of the NP itself in each case.

Adjective phrases involve the head noun of the phrase followed by any number of adjectives. For example, the phrase [baʤune tipɪs baraʔ] means "the long red skirt." This example would have a three syllable head noun [baʤune] and there would be three syllables (namely, [ti.pɪs. ba.]) intervening between the the final syllable of the noun and the final syllable of the phrase (which would be the syllable [raʔ] of [baraʔ] in this case). The points in figure 10 average over 9 such sentences, each of which had a three-syllable head noun. Of these, one had 10 syllables intervening between the final syllable of the head noun and the final syllable of the phrase, two had 5 syllables intervening, two had 3 syllables intervening, two had 1 syllable intervening, and two had the noun unmodified by an adjective. In each case, there was an F0 peak on the final syllable of the head noun and another peak on the final syllable of the phrase (which would be the last adjective in the phrase).

Stative verbs in Balinese are formed with the prefix [mɨ-]. Such verbs may be "pre-posed" out of the canonical SVO word order and appear at the beginning of the sentence. In such cases, the head of the phrase is the verb and the intervening syllables include any material in the subject noun. All of the sentences here included further material following the subject noun, such as a prepositional phrase or an adverb. For example, [mɨtakɔn iyɨ aʤaʔ gurune] means "he's asking the teacher." This example would have a three syllable stative verb [mɨtakɔn] and a two syllable subject noun [iyɨ]. The points in figure 10 average over 5 such sentences, three with 3-syllable stative verbs and two with 4-syllable stative verbs. Four of the sentences had a 2-syllable subject noun or pronoun and one had a 3-syllable subject noun. There was a peak on the final syllable of the pre-posed stative verb and another peak on the final syllable of the subject in each case.

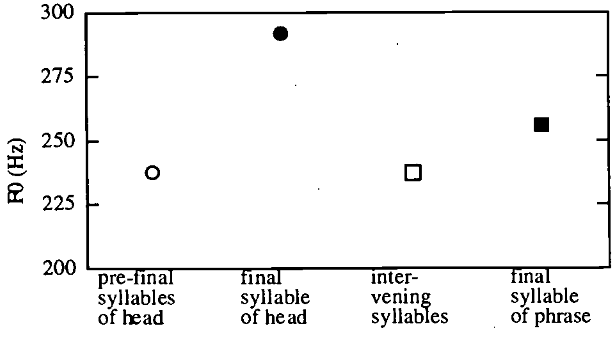


Figure 10. F0 values in more complex NPs. The sentences used here are shown in Appendix 2a, b, and c.

Figure 10 includes all three types of constructions and shows the mean F0 value of the pre-final syllables of the head (measured at the center of each vowel), the mean F0 value of the final syllable of the head (measured at the highest F0 value of the peak), the mean F0 value of all of the syllables intervening between the head and the end of the phrase (measured at the center of each vowel), and the mean F0 value of the final syllable in the subject phrase (measured at the highest F0 value of the peak). Unpaired t-tests show that the F0 values for the pre-final syllables of the head of the phrase are statistically different from the F0 values for the final syllable of the head of the phrase at the .05 level. This means that there is still an F0 peak on the final syllable of the head of the phrase, just as there was in single-word nominal subjects. Unpaired t-tests also show that the values for the syllables intervening between the head and the end of the phrase are statistically different from the values for the final syllable of the phrase at the .05 level. Again, the lower F0-value of the phrase-final peak as compared with the earlier peak may be indicative of utterance-level or phrase-level declination. Because the means for all pre-final syllables are clustered so tightly together, and because the means for all final syllables are clustered so tightly together (such that the standard error bars are not even visible here), and because the means for pre-final syllables are so well-separated from the means for final syllables, there can be said to be a consistency of alignment between the F0 peaks and syntactic positions. Thus, the accentuation is governed by the syntactic structure, although it is not dependent on syntactic categories such as "adjective" or "possessive" but rather on syntactic functions such as "head of phrase."

Interestingly, these locations match a description of F0 peak placement in Indonesian given by Laksman (1994), who writes that:

> In the realization of an NP before the verb, the highest F0 peak appears in the final syllable of the phrase. A secondary peak occurs on the last syllable of the first word. After the first word of the NP the F0 contour is generally falling until the penultimate of the final word. (Laksman, 1994, p. 128)

This seems to imply that F0 peaks in Indonesian are also aligned with the last syllable of the head of the phrase and with the last syllable of the phrase, just as they are in Balinese.

F0 peak placement in Balinese is also interesting in that it supports the cross-categorial claim of Selkirk's (1986) end-based theory of accentuation. That is, Selkirk proposes that:

> ...the relation between syntactic structure and prosodic structure above the foot and below the intonational phrase is defined in terms of the *ends* of syntactic constituents of designated types. ... The general claim I am making here is that α [the selected constituent] will be drawn from the set of categories defined in X-bar theory, and that α indicates only a level (or type) in the X-bar hierarchy, which is to say that the syntax-to-prosody structure mapping is claimed to be cross-categorial in nature. (Selkirk, 1986, p. 385)

This claim of the cross-categoriality of the syntax-to-prosody mapping would seem to be supported by the data in this section, where the F0 peak is aligned with the "same" syllable, namely, the final syllable of the head word, no matter whether it is a head noun in a genitive construction, a head noun in an adjective phrase, or a pre-posed stative verb acting as the head.

PRINCIPLES OF ACCENTUATION IN PREDICATES

The picture of accentuation in predicates is not so clear. When there is indeed an F0 peak in the predicate, its placement follows the principles outlined above. That is, there will be an F0 peak on the final syllable of the head of the verb phrase and another on the final syllable of the verb phrase itself. However, there often is not an F0 peak in the predicate at all, and the F0 contour seems to just decline steadily with no perturbations (except for segmental ones). Comparison between sentence types can be used to show the difference between predicates with and without an accent. One reason for (the appearance of) the lack of F0 accent on some predicates may be an extreme compression of the pitch-range in the sentence-final phrase. Thus, for example, passive verbs almost always show an F0 peak on their final syllable, but their agent-phrase (which follows the verb) tends to show an F0 peak only if there is another phrase following it.

Passive verbs are formed using the non-nasalized form of the verb and the valence-increasing suffix [-aŋ]. The initial segment of a verb in Balinese alternates between a nasal and a corresponding non-nasal, indicating active voice (nasal) and passive voice (non-nasal). The correspondences are shown in figure 11. The valence-increasing suffix [-aŋ] marks benefactives and causatives as well as passives– hence it marks a simple increase in the number of arguments. (In general, the resulting derived verb takes 3 arguments, including the subject.) A number of other languages, including Chuckchee and Wolof, have the same range of functions marked by a single affix (Comrie, 1985). Third-person agents are marked on the verb by the additional suffix [-ɨ]. The agent of the verb is preceded by [aʤaʔ] (a function word which can also mean "and").

passive	active
[p b]	[m]
[t d]	[n]
[ʧ ʤ s]	[ny]
[k g V]	[ŋ or ŋ+V]
[l r w m]	[ŋɨ + verb]

Figure 11. Nasal / non-nasal correspondences

The points in figure 12 average over nine 4-syllable passive verbs and two 5-syllable passive verbs. An unpaired t-test shows that the final syllable is statistically different from the pre-final syllables at the .05 level. Thus, in the predicate as well, it seems that there is an F0 peak on the final syllable of the head of the phrase.

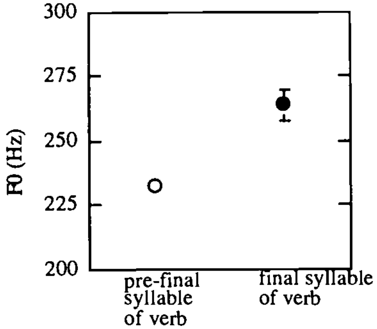


Figure 12. F0 measurements for pre-final vs. final syllables in passive verbs. The sentences used here are shown in Appendix 3a.

Comparisons between sentence-final and non-sentence-final phrases in the predicate can come from comparisons of passives whose agent-phrase occurs utterance-finally with passives whose agent-phrase is followed by other material. The comparison in this case would be between the two agent-phrases themselves– one of which is the final phrase in the sentence and one of which is not the final phrase in the sentence. For example, a prepositional phrase may occur in the middle of the sentence, allowing the agent-phrase to be sentence-final, as in the top example in figure 13. Or, the same prepositional phrase may occur at the end of the sentence after the agent, making the agent-phrase non-final, as in the lower example given in figure 13.

[lulune	intuŋanɨ	kɨ	ʧɨlabahe	aʤaʔ	uwan	tiyaɲe]
garbage-the	*was thrown away*	*to*	*river-the*	*by*	*aunt-of*	*mine*

the garbage was thrown away to the river by my aunt

[lulune	intuŋaɲɨ	aʤaʔ	uwan	tiyaɲe	kɨ	ʧɨlabahe]
garbage-the	*was thrown away*	*by*	*aunt-of*	*mine*	*to*	*river-the*

the garbage was thrown away by my aunt to the river

Figure 13. Examples of sentences with the agent phrase final in the sentence (top example) vs. non-final in the sentence (lower example). The agent phrase in underlined in each case.

The passive agent has an F0 peak on the last syllable of the phrase if the agent-phrase is non-final in the sentence. If the passive agent is the final phrase in the sentence, then it does not have an F0 peak at all, reflecting the fact that it is the final phrase and hence has a very compressed pitch range. Figure 14 shows a token of a sentence-final agent phrase and a non-final agent phrase. Only the agent phrase is shown here. There is a peak on the final syllable of the non-sentence-final agent phrase, and no peak on the final syllable of the sentence-final agent phrase.

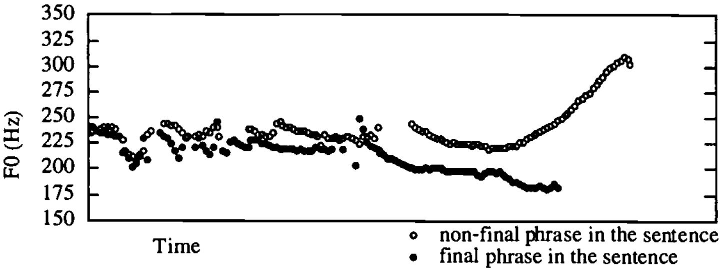


Figure 14. An F0 trace of an agent phrase which is final in the sentence overlaid on an F0 trace of an agent phrase which is non-final in the sentence. This is the agent phrase [aʤaʔ uwan tiyaɲe] "by my aunt" and only the agent phrase itself is shown here.

The points in figure 15 average over two examples with following material after the agent phrase and nine examples without following material after the agent phrase (and hence with agent phrase final in the sentence). The non-sentence-final phrases do show an F0 peak on their final syllable, while the sentence-final phrases actually show a decline in F0 on their final syllable. Unpaired t-tests show that the final syllables of both final phrases and non-final phrases are different than the pre-final syllables at the .05 level, although non-final phrases show an F0 rise while final phrases actually show an F0 fall on the final syllable. The final syllables of non-final phrases show such a large error because there are only two examples. These facts indicate that the principles of accentuation are the same in predicates as they are in subjects. Namely, there is an F0 peak on the final syllable of the head of the phrase and another F0 peak on the final syllable of the phrase itself. This effect is complicated by what appears to be utterance-final compression of F0-range, in which the sentence-final phrase does not display any of the expected F0 peaks.

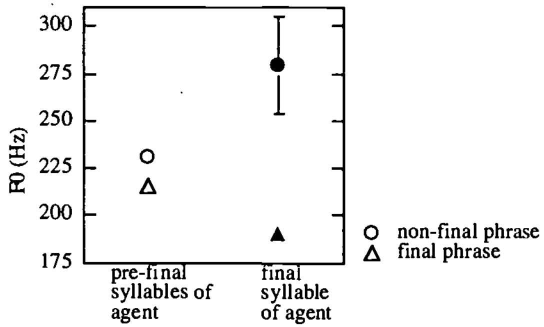


Figure 15. Passive agents that are final in the sentence vs. non-final in the sentence. The sentences used here are shown in Appendix 3a.

A similar comparison to the one just made between sentence-final and non-final agent phrases can be made between ditransitive verbs marked with the valence-increasing suffix [-aŋ] (and hence taking 2 objects) and those without it (which take either 0 or 1 object). For example, a comparison can be made between the sentences in figure 16.

[tiyaŋ	<u>mili</u>	sɨpatu	baru]
I	*bought*	*shoes*	*new*

I bought new shoes

[tiyaŋ	<u>milian</u>	adɪn	tiyaŋe	sɨpatu	baru]
I	*bought-for*	*brother-of*	*mine*	*shoes*	*new*

I bought new shoes for my brother.

Figure 16. Example sentences with a simple verb vs. with a valence-increasing verb.

In these cases, since the unsuffixed verbs are part of the sentence-final phrase, they show no F0 peak. The suffixed verbs, on the other hand, can have a benefactive interpretation with two objects, so the verb is not part of the sentence-final phrase and hence can have an F0 peak on it. The effect shown here is slightly confounded by the immediately preceding peak in the final syllable of the subject phrase. The presence of that peak means that the first syllable of the verb is higher than might be expected, since it is in the decline from the previous peak. The points in figure 17 average over 10 examples of simple verbs and 28 examples of suffixed, valence-increasing verbs. The unsuffixed forms examined here are all disyllabic and the suffixed forms are all trisyllabic. Unpaired t-tests show that although the prefinal syllables of suffixed forms are different from the final syllables of suffixed forms at the .05 level (because there is a peak present), the prefinal syllables of the unsuffixed forms were not statistically different from the final syllables of the unsuffixed forms at the .05 level (because the final syllable is part of a gradual decline towards the end of the sentence).

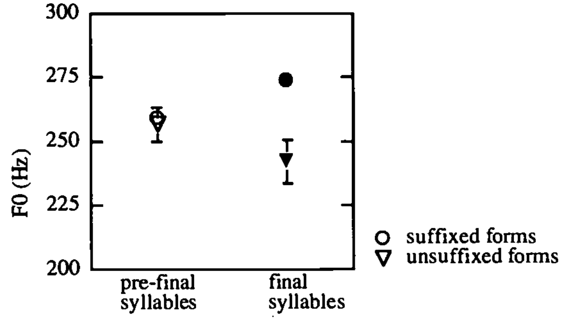


Figure 17. Pre-final vs. final syllables of simple verbs vs. valence-increasing verbs. The sentences used here are shown in appendix 3b(i) and (ii).

The picture of accentuation in the first object of valence-increased verbs is not very clear at all. It seems that if there is an F0 peak present, its location is syntactically determined. However, sometimes there is no accent at all. Moreover, whether there is an accent present or not can vary even from one repetition to another of the same sentence. The first object phrase from two tokens of the same sentence are shown in figure 18. In one token, there is an F0 peak on the final syllable of the phrase and in the other, there is no peak.

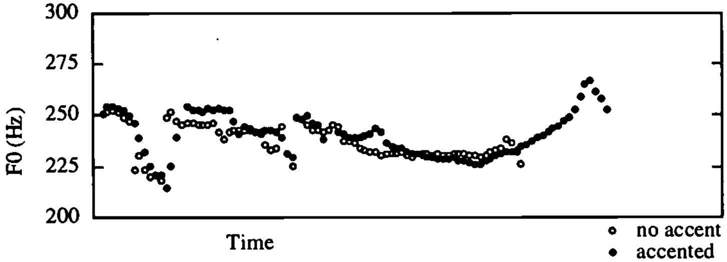


Figure 18. Two tokens of the same utterance, one with and one without accent. (Only the first object itself is shown here.) [tiyaŋ miliaŋ adın tiyaŋe sipatu putıh mitali barah] "I bought new white shoes laced with red for my brother."

If there is an accent present in the first object of a double-object construction, it follows the same principles of accentuation described above, that is, it falls on the last syllable of the first object NP. The points in figure 19 average over 11 tokens, 6 with 3-syllable simple nouns as object 1 ([panaʔne] "his/her children" or [bayine] "his/her baby") and 5 with more complex NPs. Two of the more complex NP tokens were [adın tiyaŋe] "my brother," two were [buku barune] "the red book," and one was [panaʔne muani abısıh] "his/her only son." An unpaired t-test shows that pre-final syllables are different from final syllables at the .05 level. The fact that all of the pre-final syllables of the object have a lower F0 and that there is only ever one peak, even in more complex NPs, may indicate that this object phrase is being phrased together with the verb, and that the verb is the head of the phrase.

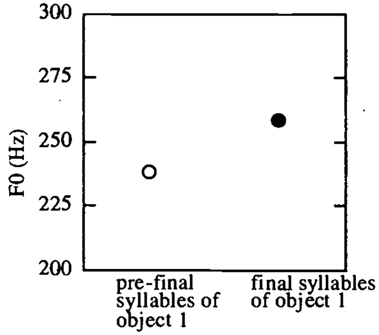


Figure 19. Accented first object in double-object constructions. The sentences used here are shown in appendix 3b(ii).

An explanation for the lack of accent in the cases in which there is no accent in the first object may be a type of "de-phrasing," where a potential accent is not realized. This view is supported by data from spontaneous narratives, where there often is not an F0 peak where an accent might be expected from the principles described above. Thus, it seems that the principles discussed above describe where an accent can be placed if one is indeed to be placed in the sentence, but they by no means mandate that there must be an accent present in that location. Further investigation may bring out factors which determine the amount of "de-phrasing" vs. accent realization. Such factors might be expected to include speech rate or speech style.

In the second object of double object constructions, the only time an accent was found among these data was in case there was a stative verb/predicative adjective present in the predicate. In Balinese, the predicative adjectives pattern with the stative verbs in some respects. Predicative adjectives can take verbal suffixes such as the valence-increasing suffix [-aŋ] to form a causative with concomitant nasalization of the first consonant (indicating active voice) and they can also be marked for tense or aspect. Thus, Balinese would fall into the "adjectival-verb" class of languages as opposed to the "adjectival-noun" class of languages (in the typology suggested by Schachter (1985)), although the adjectives would form a subset of the verbs since they can also be used attributively or with comparative or superlative marking. In any event, the only cases found in the corpus where there was an F0 peak on the second object of double object constructions was before stative verb/predicative adjectives. For example, in noun phrases such as the following there could be an F0 peak on the last syllable of the attributive adjective "big," before the stative verbs/predicative adjectives indicating "3-storied" and "metal-fenced."

umah	gɨde	mɨtiŋkat	tɨlu	mɨpagɨhan	bɨsi
house	*big*	*storied*	*three*	*fenced*	*metal*

a big 3-storied metal-fenced house

Figure 20. Double-object construction which could have an accent on the second object.

There could be an F0 peak on the last syllable of "big" whether it was followed by either one of the two predicative adjectives listed above or by both. In these cases, it may be that the stative verb/predicative adjectives are being set off in their own phrases. Again, there are cases where two tokens of the same utterance show different accentuation, as in figure 21. One token shows an F0 peak on the last syllable of the adjective "big," while the other token of the same phrase shows no F0 peaks at all within the second object phrase.

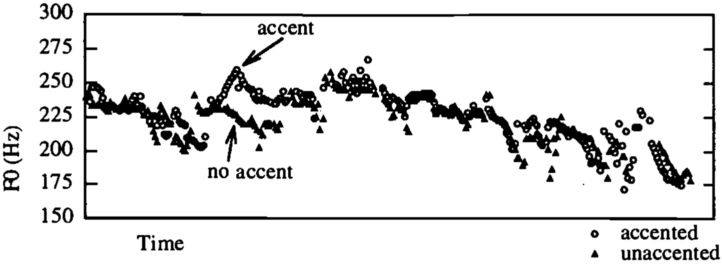
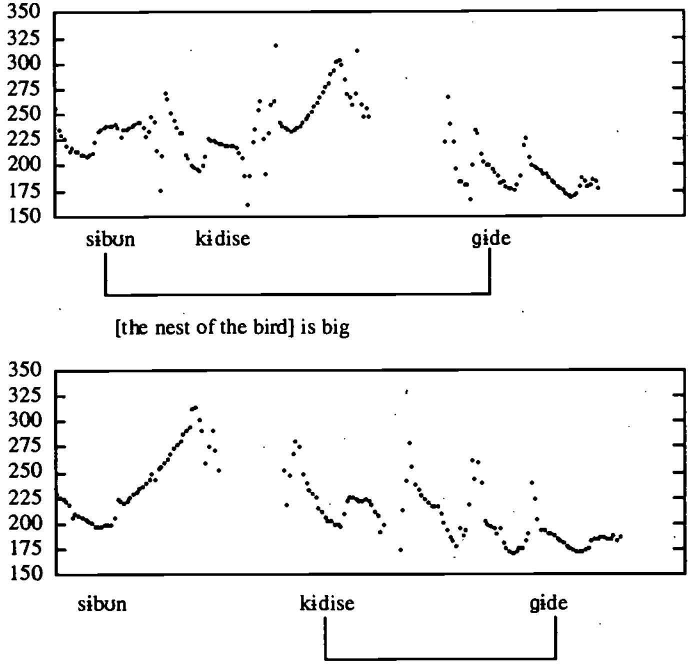


Figure 21. F0 traces of two tokens of the same second object in double-object constructions. One of these repetitions has an accent on the final syllable of the word [gɨde] "big" and the other does not have an accent in that location, or anywhere. The sentence shown here is [iyɨ ŋaenaŋ panaʔne umah gɨde mɨtiŋkat tɨlu mɨpagɨhan bɨsi] "he builds his child a big three-storied, metal-fenced house" although only the underlined part (the second object itself) is shown here.

Thus, these cases seem to be showing yet again the interaction between having an F0 peak on the final syllables of heads of phrases and the last syllables of phrases and reducing the F0 range sentence-finally. Again, the principles for accent-placement seem to legislate where an accent can go should an accent be placed, but they do not mandate that an accent must be present, and there seems to be some freedom involved as to whether an accent should be placed or not. Further research would be needed to determine whether accent-placement vs. no accent in such cases has more subtle pragmatic meanings that could possibly be teased out of the sentence.

PHRASAL "CONTRASTS"

It is possible to disambiguate two potentially ambiguous sentences using the principles of accentuation described above. For example, the sentence [sɨbun kɨdise gɨde] "nest-of bird-the big" can have two interpretations. One interpretation is "the nest of the bird is big" if "big" is understood to modify "the nest of the bird" (since there can be adjective constructions in Balinese without a copula). This interpretation can be facilitated by a peak on "bird-the," which phrases "nest-of" and "bird-the" together and leaves "big" modifying the whole preceding phrase. Another interpretation is "the nest of the big bird" if "big" is understood to modify "bird." This interpretation is facilitated by a peak on "nest-of," setting it in its own phrase apart from "bird-the" and "big" which are then phrased together. These two phrasings are shown in figure 22. Thus, although lexical contrasts through accentuation are not possible in Balinese because lexical accentuation is not distinctive, phrasal contrasts through accentuation are possible because the accentuation occurs at the phrasal level, and so can serve to distinguish different phrasings.



[the nest of the bird] is big

the nest of [the big bird]

Figure 22. F0 traces showing examples of disambiguating via phrasing. The two sentences shown here are lexically identical, but prosodically differentiated. Thus, the adjective [gide] "big" is modifying either the noun [sibun] "the nest of" or the noun [kidise] "the bird."

PRINCIPLES OF ACCENTUATION IN CLAUSES

Relative clauses in Balinese are formed with the relativizer [ane] and the relative clause embedded after the subject NP. In such constructions, there is the usual accentuation in the subject phrase with a rise on the final syllable of the head and another rise on the final syllable of the phrase, just before the relativizer. The points in figure 23 average over four tokens with relative clauses. The large error in the final syllables comes from averaging over only 4 F0 values, while the pre-final syllable values come from an average over many more syllables.

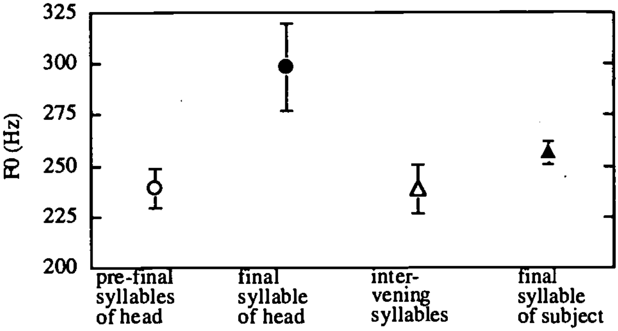


Figure 23. F0 values in complex NPs modified by relative clauses. These values are from the subject itself. The sentences used here are shown in Appendix 4a.

Within the relative clause itself, there is also an F0 rise on the final syllable, with all of the pre-final syllables showing a consistently low F0 value. Unpaired t-tests show that the pre-final syllables are statistically different from the final syllables at the .05 level. So again, the accentuation is serving a delimitative purpose here, by setting off the relative clause from the rest of the sentence.

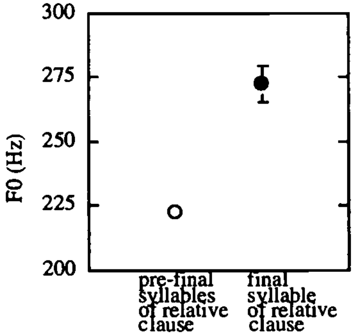


Figure 24. Pre-final syllables within the relative clause vs. the final syllable of the relative clause. The sentences used here are shown in appendix 4a.

In conditionals, the "if" clause has an F0 peak on the last syllable of the word "if" and on the last syllable of the clause before the "consequence" clause, which in the tokens shown here is an imperative. The "if" clauses used here varied in terms of syntactic structure. The points in figure 23 average over 6 examples. Since there is no peak on the NP within the clause, this seems to indicate that the word "if" is itself the head of the clause.

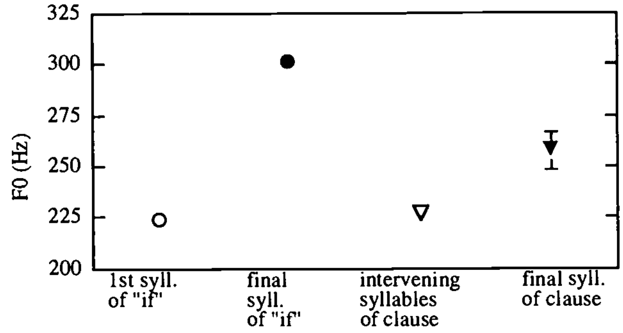


Figure 25. F0 values of conditionals.The sentences used are shown in Appendix 4.

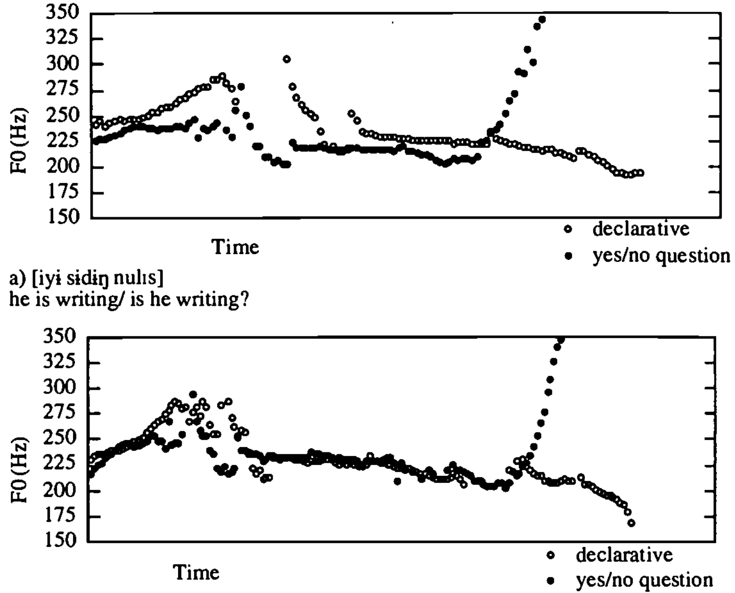
Thus, accentuation works the same way in relative clauses and in conditionals as it does in the phrases described above. There is an F0 peak on the final syllable of the head and another F0 peak on the final syllable of the clause itself.

SENTENCE-LEVEL INTONATION

Accentuation in Balinese, as shown above, is syntactically governed. However, F0 may also be used to indicate the pragmatic function of a sentence. In Balinese it is possible to form pragmatically different utterances using identical lexical items, via intonation. For example, yes/no interrogatives may have the same string of words as indicative sentences, but may be differentiated by F0 contour. The yes/no questions in Balinese differ from declaratives in two ways. First, there is a sharp rise on the final syllable of the yes/no question. It does not matter how long the utterance is, this rise will coincide with the last syllable of the utterance only. Second, the F0 peak usually present on the last syllable of the subject in 1-word nominal subjects is missing in yes/no questions. Both of these phenomena are shown in the figure 26, in which two examples are given of lexically identical sentences differentiated by intonation.

This type of yes/no question formation, in which the interrogative status of the sentence is conveyed not by lexical differences but by the F0 pattern, is seen in other languages as well. For example, Bolinger (1978) surveys languages which have rising terminals in yes/no questions and cites many examples.

Thus, F0 may also serve pragmatic functions in Balinese and not just delimitative functions. In fact, the pragmatic functions seem to override the syntactically governed accentuation, as suggested by the lack of accent on the subjects in yes/no questions.



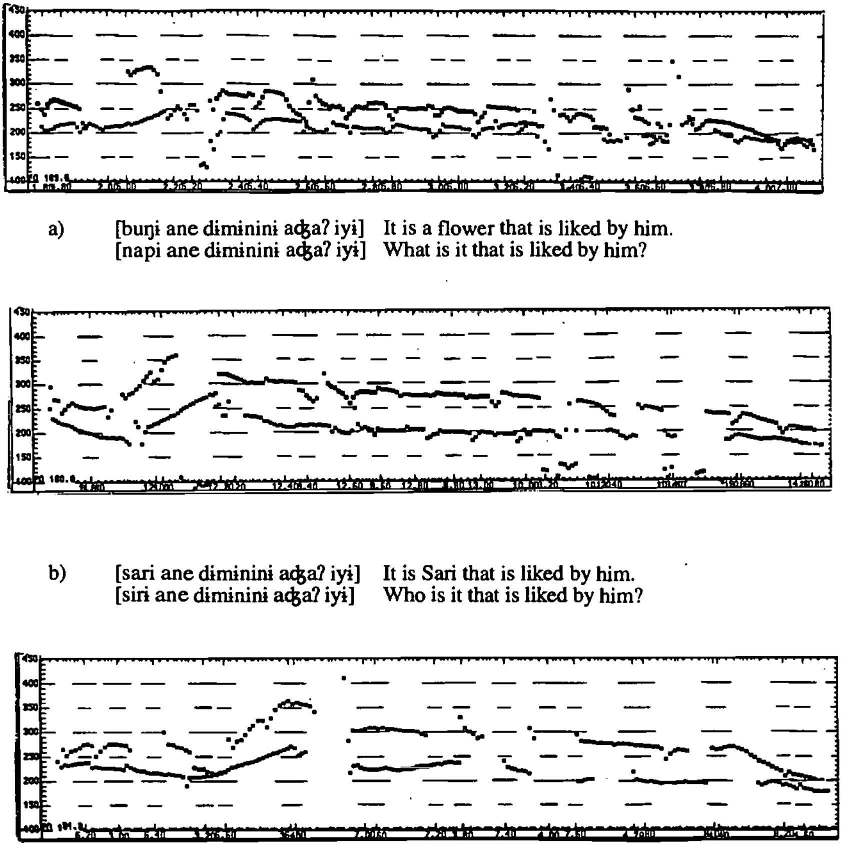
a) [iyɨ sɨdɨŋ nulɪs]
he is writing/ is he writing?

b) [iyɨ sɨdɨŋ nulɪs surat]
he is writing a letter/ is he writing a letter?

Figure 26. F0 traces of declarative sentences overlaid on F0 traces of yes/no interrogatives.

In Balinese, it is also possible to form "wh-questions" that are very similar to declaratives, with a "wh-word" substituted for a noun. In these cases, the "wh-question" has a higher F0 range than the declarative and the "wh-word" itself seems to have a steeper F0 rise than the noun it replaces, as seen in figure 27.

The F0 patterns seen in "wh-questions" in Balinese resemble those seen in Tokyo Japanese (Maekawa, 1991) and in Korean (Jun and Oh, 1994). In Tokyo Japanese, it is possible to form two lexically similar sentences where one is a "wh-question" and one is an indefinite.The "wh-questions" are prosodically different from the indefinites in two ways– the F0 peak on the "wh-word" is more salient and the "wh-sentence" consists of one intermediate phrase while the indefinite contains a prosodic boundary and hence has two intermediate phrases. In Korean, there is also a prosodic difference between "wh-words" as used in "wh-questions," in incredulity readings, and as indefinite pronouns in yes/no questions. The three question types are distinguished by boundary tones and phrasing, with the "wh-questions" and incredulity questions in one accentual phrase and the pitch being higher in incredulity questions than in yes/no questions and higher in yes/no questions than in "wh-questions." So the patterns of F0 expansion throughout the sentence in Balinese and particularly on the "wh-word" itself are reminiscent of the patterns seen in other languages.



a) [buŋɨ ane diminini aʤaʔ iyɨ] It is a flower that is liked by him.
[napi ane diminini aʤaʔ iyɨ] What is it that is liked by him?

b) [sari ane diminini aʤaʔ iyɨ] It is Sari that is liked by him.
[siri ane diminini aʤaʔ iyɨ] Who is it that is liked by him?

c) [bɪn mani iyɨ kɨ sikɔlahan] Tomorrow he will go to school.
[bɪn pidan iyɨ kɨ sikɔlahan] When will he go to school?

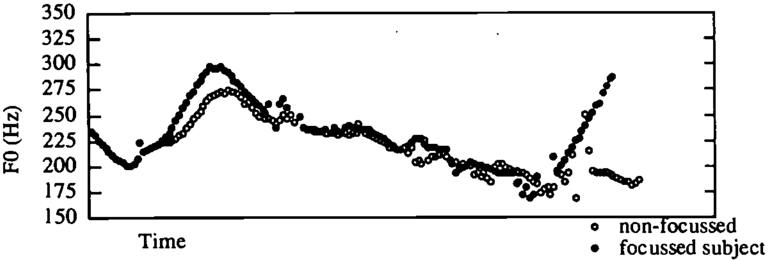
Figure 27. F0 traces of declarative sentences overlaid on F0 traces of wh-questions.

Another pragmatic use of F0 in Balinese is to focus a particular element in the sentence. This type of situation can be elicited through a set of questions which all prompt the "same" answer (lexically speaking) but which require the speaker to contradict a different element of the question in each answer. (Two sets of similar examples are listed in Appendix 5.) For example, one answer might be:

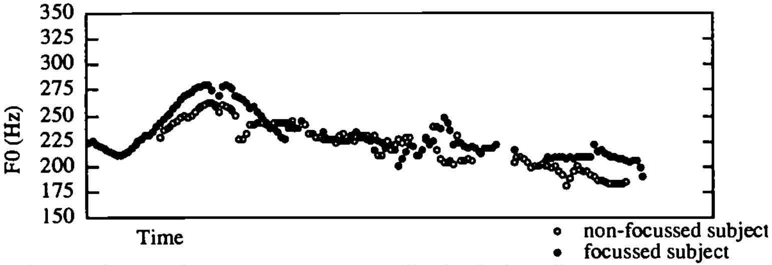
[tiyaŋ lakar ŋumbah baʤu] "I will wash clothes."

Various questions can be asked which prompt the "same" sentence as a response. For example:

a) [lakar ŋuʤaŋ ?] "What will you do?" should elicit a non-focussed reply.
b) [iyɨ lakar ŋumbas baʤu ?] "Will she wash clothes?" should elicit a reply with narrow focus on the subject.
c) [lakar ninʔ ?] "Will you iron?" should elicit a reply with narrow focus on the verb.
d) [lakar ŋumbah siprai] "Will you wash the spread?" should elicit a reply with narrow focus on the object.
e) [sampʊn ŋumbah baʤu ?] "Did you already wash the clothes?" should elicit a reply with narrow focus on the tense marking.

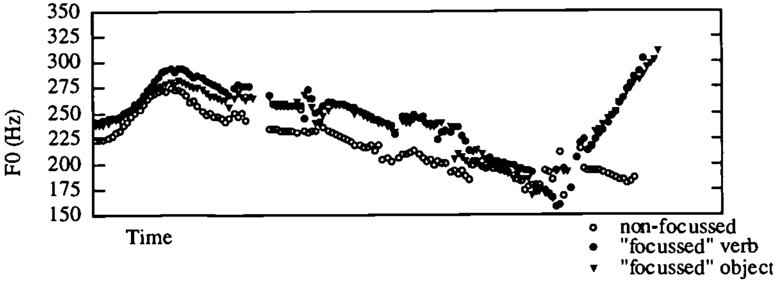


[tiyaŋ lakar ŋumbah baʤu] "I will wash clothes" as answers to:
a) [lakar ŋuʤaŋ?] "What will you do?"
b) [iyɨ lakar numbah baʤu?] "Will s/he wash the clothes?"



[iyɨ lakar nɨʧɛt umah] "She will paint the house" as answers to:
a) [iyɨ lakar ŋuʤan ?] "What will she do?"
b) [ragane lakar ŋɨʧɛt umah?] "Will you paint the house?"

Figure 28. F0 traces of sentences with focussed subjects overlaid on F0 traces of non-focussed sentences.



[tiyaŋ lakar ŋumbah badʤu] "I will wash clothes" as answers to:
a) [lakar ŋuʤaŋ?] "What will you do?"
b) [lakar nɨrɪ??] "Will you iron?"
c) [lakar ŋumbah siprai?] "Will you wash the spread?"

Figure 29. F0 traces of sentences with "focussed" verb or object overlaid on an F0 trace of a non-focussed sentence.

The question is whether the speaker can prosodically focus certain elements of the sentence by putting the F0 accent in a higher pitch range than normal. At first glance, some elements of the sentence did appear to have a higher F0 when that element was in narrow focus. In order to test the salience of the alleged prosodic focus in various locations in the sentence, a listening experiment was performed several weeks after the recording. The answer-sentences which had been elicited earlier were played back to the speaker in random order and she was asked what would be an appropriate question that would elicit each sentence as an answer. She correctly proposed a general question for non-focussed sentences, a question about the subject for sentences with the subject under focus, and a question about the tense marking for the sentences with the tense marking under focus, but the sentences with the allegedly focussed object and the sentences with the allegedly focussed verb sounded the same to her (and so she did not consistently propose the correct question). It seems like elements in the sentence which are accentable anyhow according to the principles described earlier can be prosodically focussed. That is, in the sentence above, the subject and the tense marking can be prosodically focussed but not the verb or the object, which are in the sentence-final phrase and hence usually do not have an accent at all. The examples in figure 28 and 29 compare a general, non-focussed sentence with a sentence with narow focus on the subject, on the verb, and on the object. (The sentence-final syllable may show a continuation rise, which may not be relevant to the focus issue at all.) The sentences with narrow focus on the subject (the two displays in figure 28) do show a higher peak on the final syllable of the subject than the declarative sentences have, but the sentences with the narrow focus on the verb and on the object (the display in figure 29) are quite similar to the non-focussed sentence.

Interestingly, the un-focus-able elements, such as the object, can be syntactically focussed by being topicalized and appearing at the beginning of the sentence, as in [tin, urutan ane lakar gaenɨ] "no, it is sausage that will be made" or [tin, umahne ane lakar ʧetɨ] "no, it is the house that will be painted." So a tentative conclusion on prosodic focus in Balinese is that focus can enhance an F0 peak that would be expected to be present in any case, but it cannot place an F0 peak in an "unaccentable" location in the sentence. Some such "unaccentable" locations, though, can be syntactically topicalized, thus allowing them to be focussed by position rather than by prosody (although it might be argued that the motivation for the focus-by-position option is in fact the prosodic restrictions on "accentable" vs. "unaccentable" positions).

CONCLUSION

In conclusion, in Balinese there is no lexical level accentuation. At the phrasal level, there is an accent on the last syllable of the head of the phrase and on the last syllable of the phrase, although this effect is confounded in predicates by a sentence-final F0 compression. The accentuation is cross-categorial, occurring on any phrasal head and on any final syllable, regardless of their syntactic categories. There is also pragmatic use of F0 made in Balinese, including yes-no question formation, wh-question formation, and focus. Thus, Balinese shows no word-level accentuation but completely regular phrasal use of accentuation, as indicated by the consistency of alignment between F0 peaks and particular syntactic positions.

APPENDIX 1: sentences with single-word subjects

a) 1-word nominal subjects:

yɛhe	sampʊn	mɨluwab
water-the	*already*	*boils*

the water's already boiling

iyɨ	nyaɪt	bantɨn	aʤaʔ	ŋae	ʤaʤɨ
she	*makes*	*crafts*	*and*	*makes*	*cookies*

she makes crafts for offereings and makes cookies

tiyaŋ	mambʊh	tuni	sɨmɨŋan	kɨranɨ	bɔʔ	tiyaɲe
I	*wash*	*earlier*		*because*	*hair*	*mine*
sampʊn	daki					
already	*dirty*					

I washed my hair earlier this morning because my hair was already dirty.

tiyaŋ	ŋɨdupaŋ	lampu	apaŋ	galaŋ	umahe
I	*turn on*	*lamp*	*for*	*lighting*	*house-the*

I turn on the lamp for lighting the house.

limane	ʤɨpɨs	ʤɨlanan
hand-his	*caught*	*door*

his hand is caught in the door

sudiri	mɨʤuʤʊʔ	di	pɨsareane
sudiri	*stands*	*on/at*	*bed-the*

Sudiri stand on the bed

paʤɨɲe	bɨnahinɨ	aʤaʔ	ibʊn	pɨsagan	tiyaɲe
umbrella-the	*was fixed*	*by*	*mother-of*	*neighbor-of*	*mine*

the umbrella was fixed by my neighbor's mother

kurɨnane	durʊŋ	lulʊs	di	sɨkɔlahan
wife-his	*not yet*	*graduated*	*from*	*school*

his wife has not yet graduated from school

gɨgelane	mɨʤuʤʊʔ	di	kursine
boyfriend-her	*stands*	*on*	*chair-the*

her boyfriend stands on the chair.

sɨpedane	pɨsilihaɲɨ	aʤaʔ	gɨgelane
bicycle-the	*was lent*	*by*	*boyfriend-her*

the bicycle was lent by her boyfriend

kɨponaane	sampʊn	gɨde gɨde
niece-her	*already*	*grown-up*

her niece is already grown up

pɨrabotane	luwʊŋluwʊŋ	gati
furniture-his	*good*	*very*

his furniture is very good

matɛmatɪkane	maŋkɪn	ŋantʃan	sukɨh
math	*now*	*become/more*	*difficult*

math now becomes difficult

b) conjoined nominal subjects:

bapane	adʒaʔ	ibune	niŋɔɪn	iyɨ	di	rumah	sakɪt
father-his	*and*	*mother-his*	*visit*	*him*	*at*	*house*	*sick*

his father and his mother visit him at the hospital

bapane	napi	bɨline	ane	lakar	ŋatɪhɪn	iyɨ	kɨ	dɔktɨr
father-his	*or*	*brother-his*	*that*	*will*	*drop off*	*him*	*at*	*doctor*

it is his father or his brother who will drop him off to the doctor

iyɨ	adʒaʔ	timpalne	kɨ pɨkɨn	barɪŋbarɪŋ
he	*and*	*friend-his*	*to market*	*together*

he and his friend went to market together

APPENDIX 2: sentences with more complex subjects

a) genitive constructions:

ibʊn	tiyaɲe	mɨlandʒaran	kɨ	umah	timpalne
mother-of	*mine*	*travels to*	*house*	*friend-hers*	

my mother travels to her friend's house

adɪn	tiyaɲe	mɨumah umahan	di	bitɛn	medʒɨne
brother-of	*mine*	*plays house*	*on/at*	*bottom*	*table-the*

my brother plays house under the table

adɪn	tiyaɲe	mɨumah umahan	di	duwʊr	punyan
brother-of	*mine*	*plays house*	*at*	*top*	*tree*
kayune	tɨgɨh				
wood-the	*tall*				

my brother plays house at the top of the tall tree

adɪn	tiyaɲe	sampʊn	suwʊd	masʊʔ	di	bandʊŋ
brother-of	*mine*	*already*	*finished*	*studying*	*at*	*bandung*

my brother already finished studying at bandung

miyɔɲ	tiyaɲe	adʒaʔ	tʃitʃɪŋ	timpal	tiyaŋe	mɨkɨrah	dɔgɛn	gaenɨ
cat	*mine*	*and*	*dog*	*friend*	*mine*	*fight*	*always*	*make*

my cat and my friend's dog always fight

duwan	tiyaɲe	bisi`	mɨnɛʔ	punyan	nyʊh
aunt-of	*mine*	*can*	*climb*	*tree*	*coconut*

my aunt can climb a coconut tree

umah	tiyaɲe	adʒaʔ	kantɔr	pɔse	mɨpaɨʔan
house	*mine*	*and*	*office*	*post*	*close*

my house and the post office are close

misanan		tiyaɲe	tɨn	dɨmɨn	ŋango	rɔʔ	kɨranɨ	iyɨ
cousin		*mine*	*not*	*like*	*wear*	*skirt*	*because*	*she*
masɨ	kɛwɨh	gati	lamɨn	iyɨ	ŋango	rɔʔ		
feels	*awkward*	*very*	*if*	*she*	*wears*	*skirt*		

my cousin doesn't like to wear skirts because she feels awkward when she does

pisagan	tiyaɲe	nɨgaʔ	di	kursine
neighbor	*mine*	*sits*	*at/on*	*chair-the*

my neighbor sits on the chair

kɨponaan	tiyaɲe	sampʊn	bisɨ	mɨʤalan
niece-of	*mine*	*already*	*can*	*walk*

my niece already can walk

warnan	paʤɨŋ	iyɨne	baraʔ	aʤaʔ	putɪh
color-of	*umbrella-of*	*his-the*	*red*	*and*	*white*

the color of his umbrella is red with white

bapan	timpal	tiyaɲe	nyumunɪn	ŋae	bale	bɪn	tɨlʊn
father-of	*friend*	*mine*	*begins*	*make*	*gaz.*	*fut.*	*three*

my friend's father begins to make a traditional gazebo in three days

ibu	timpal tiyaɲe	bisɨ	ŋae	baʤu
mother	*friend mine*	*can*	*make*	*clothes*

my friend's mother can make clothes

ibʊn	pisagan	tiyaɲe	mɨsilihaŋ	misanan
mother-of	*neighbor-of*	*mine*	*lent*	*cousin-of...*

my neighbor's mother lent my friend's father's cousin...

b) adjectives:

baʤune	(((tipɪs)	baraʔ)	lambɪh)	pantiɲɨ	aʤaʔ	i meme
skirt-the	*long*	*red*	*thin*	*was washed*	*by*	*mother*

the (((long) red) thin) skirt was washed by mother

ʧiʧiɲɨ	((((gɨde)	badɨŋ)	galaʔ)	mikupiŋ	dawɨ)	nyɨgʊt	anaʔe
dog-the	*big*	*black*	*mean*	*eared-*	*long*	*bit*	*person-*

the

tuwɨ	ɲidɪh idɪh
old	*beggar*

the ((((big) black) mean) long-eared) dog bit the old beggar-person

c) pre-posed stative verbs:

mɨtakɔn	iyɨ	aʤaʔ	gurune
asks	*he*	*with*	*teacher-the*

he's asking the teacher

mɨblanʤɨ	ibu	sɨbilaŋ	sɨmiŋan	kɨ	pɨkin	ane
shops	*mother*	*every*	*morning*	*at*	*market*	*that*

paɨʔ	aʤaʔ	umahe
close	*with*	*home-the*

mother shops every morning at the market that's close to the house

mɨʤalan	iyɨ	di	sisɪn	pasihe
walks	*he*	*at*	*edge*	*ocean-the*

he's walking at the beach

mɨkaʊʔan	iyɨ	kɨras kɨras
shouts	*he*	*loudly*

he shouts loudly

mɨgulɪʔan	batune	marɨ	inʤɨʔ	tiyaŋ
rolls	*stone-the*	*just/as*	*stepped*	*I*

the stone rolled when I stepped on it

APPENDIX 3: Accent in Predicates

a) passive verbs:

lulune	intuŋaŋɨ	aʤaʔ	uwan	tiyaɲe	ki	ʧɨlabahe
garbage-the	*was thrown away*	*by*	*aunt-of*	*mine*	*to*	*river-the*

the garbage was thrown away by my aunt to the river

umahne	kɨdasinɨ	aʤaʔ	pɨmbantune	sɨbilaŋ	wayi
house-the	*was cleaned*	*by*	*maid-the*	*every*	*day*

the house was cleaned by the maid every day

lulune	intuŋaŋɨ	kɨ	ʧɨlabahe	aʤaʔ	uwan	tiyaɲe
garbage-the	*was thrown away*	*to*	*river-the*	*by*	*aunt-of*	*mine*

the garbage was thrown away to the river by my aunt

umahne	kɨdasinɨ	sɨbilaŋ	wayi	aʤaʔ	pɨmbantune
house-the	*was cleaned*	*every*	*day*	*by*	*maid-the*

the house was cleaned every day by the maid

panaʔne	pɨlayibaŋɨ	kɨ	bulɛlɛŋ	aʤaʔ	bapane
child-the	*was whisked away*	*to*	*Buleleng*	*by*	*father-his*

the child was whisked away to Buleleng by his father

sibɯn	kɨdise	uwuŋaŋɨ	aʤaʔ	adın	tiyaɲe
nest-of	*bird-the*	*was destroyed*	*by*	*little sibling-of*	*mine*

the bird's nest was destroyed by my little sibling

ʤɨndelɨ	kaʧane	bɨlahaŋɨ	aʤaʔ	rarene
window	*glass-the*	*was broken*	*by*	*kids-the*

the glass window was broken by the children

paʤɨɲe	bɨnahinɨ	aʤaʔ	ibɯn	pisagan	tiyaɲe
umbrella-the	*was fixed*	*by*	*mother-of*	*neighbor-of*	*mine*

the umbrella was fixed by the mother of my neighbor

rɔʔne	bɨnahinɨ	aʤaʔ	ibɯn	tiyaɲe
skirt-the	*was fixed*	*by*	*mother-of*	*mine*

the skirt was fixed by my mother

kipas	aɲine	pɨsilihaŋɨ	aʤaʔ	ibɯn	pisagan	tiyaɲe
fan	*wind-the*	*was lent*	*by*	*mother-of*	*neighbor-of*	*mine*

the electric fan was lent by the mother of my neighbor

kɨ	ʧɨlabahe	lulune	intuŋaŋɨ	aʤaʔ	uwan	tiyaɲe
to	*river-the*	*garbabe-the*	*was thrown away*	*by*	*aunt-of*	*mine*

the garbage was thrown away to the river by my aunt

b) double object constructions:

i) simple verbs:

(2 tokens each:)

tiyaŋ	maʧɨ	maʤalah
I	*read*	*magazine*

I'm reading a magazine.

iyɨ	nɔŋɔs	di	umah	tiyaɲe
he	*stays*	*at*	*house*	*mine*

He stays at my house.

tiyaŋ	mɨli	sɨpatu	baru
I	*bought*	*shoes*	*new*

I bought new shoes

(1 token each:)

iyɨ	mati
he	*died*

he died

iyɨ	ŋae	umah
he	*builds*	*house*

He builds a house.

bayine	sirɨp	di	sirɨpane
baby-the	*sleeps*	*in*	*bed-poss.*

the baby sleeps in its bed

ii) valence-increasing verbs:

tiyaŋ	ŋadipaŋ	iyɨ	buku
I	*sell-for*	*him*	*book*

I'm selling a book for him

(2 tokens)

tiyaŋ	mɨliaŋ	adɪn	tiyaɲe	sɨpatu	baru
I	*bought-for*	*brother-of*	*mine*	*shoes*	*new*

I bought new shoes for my brother

(2 tokens)

tiyaŋ	mɨliaŋ	adɪn	tiyaɲe	sɨpatu	putɪh	mɨtali	barah
I	*bought-for*	*brother-of*	*mine*	*shoes*	*white*	*laced*	*red*

I bought new white shoes laced with red for my brother.

(2 tokens)

iyɨ	ŋaenaŋ	panaʔne	umah
he	*builds-for*	*children-poss.*	*house*

he builds a house for his children

(2 tokens)

iyɨ	ŋaenaŋ	panaʔne	muani	abɨsɪʔ	umah
he	*builds-for*	*child-poss.*	*male*	*only*	*house*

he builds a house for his only son.

iyɨ	ŋaenaŋ	panaʔne	umah	gɨde	mɨpagɨhan	bɨsi
he	*builds-for*	*child-poss.*	*house*	*big*	*fenced*	*metal*

he builds a big house with a metal fence for his child

iyɨ	ŋaenaŋ	panaʔne	umah	gɨde	mɨtiŋkat	tɨlu
he	*builds-for*	*child-poss.*	*house*	*big*	*storied*	*3*

he builds a big 3-storied house for his child

iyɨ	ŋaenaŋ	panaʔne	umah	gɨde
he	*builds-for*	*child-poss.*	*house*	*big*

he build for his child a big house

(2 tokens)

iyɨ	ŋaenaŋ	panaʔne	muani	abɪsɪh	umah	gɨde	mɨtiŋkat
he	*builds-for*	*child-poss.*	*male*	*only*	*house*	*big*	*storied*
tɨlu	mɨpagɨhan	bɨsi					
3	*fenced*	*metal*					

he builds for his only son a big 3-storied house with a metal fence

iyɨ	ŋaenaŋ	panaʔne	umah	gɨde	mɨtiŋkat	tɨlu
he	*builds-for*	*child-poss.*	*house*	*big*	*storied*	*3*
mɨpagɨhan	bɨsi					
fenced	*metal*					

he builds for his child a big 3-storied house with a metal fence

(2 tokens)

tiyaŋ	ŋɛdɛɲaŋ	buku	barune	aʤaʔ	iyɨ
I	*show-to*	*book*	*new-the*	*to*	*him*

I show the new book to him.

(2 tokens)

tiyaŋ	ŋɛdɛɲaŋ	bukune	tɨbɨl	mɨkulɪt	baraʔ	aʤaʔ	iyɨ
I	*show-to*	*book-the*	*thick*	*covered*	*red*	*to*	*him*

I show the thick book covered in red to him.

(2 tokens)

tiyaŋ	ŋɛdɛŋaŋ	bukune	barune	aʤaʔ	iyɨ	mɨkɨʤaŋ
I	*show-to*	*book-the*	*new*	*to*	*them*	

I show the new book to them

(2 tokens)

tiyaŋ	ŋɛdɛŋaŋ	bukune	tibɨl	mɨkulɪt	baraʔ	aʤaʔ
I	*show-to*	*book-the*	*thick*	*covered*	*red*	*to*
iyɨ	mɨkɨʤaŋ					
them						

I show the thick book covered in red to them.

iyɨ	nyiɾipaŋ	bayine	di	siɾipane
she	*sleep-causative*	*baby-her*	*in*	*bed-the*

She puts her baby to sleep in the bed.

iyɨ	nɔŋɔsaŋ	panaʔne	di	umah	tiyaŋe
he	*house-causative*	*children-poss.*	*at*	*house*	*mine*

He houses his children at my house.

iyɨ	nɔŋɔsaŋ	panaʔne	lʊh	ʤigɛg	mɨkulɪt	kɨdas
he	*house-causative*	*child-poss.*	*female*	*beautiful*	*skinned*	*clean*
di	umah	tiyaŋe				
at	*house*	*mine*				

He houses his beautiful daughter with clean skin at my house.

iyɨ	nɔŋɔsaŋ	panaʔne	di	umah	gɨde
he	*house-causative*	*children-poss.*	*at*	*house*	*big*
gɨlah	pamane				
owned-by	*uncle-his*				

He houses his children at the big house owned by his uncle.

APPENDIX 4: Clauses

a) relative clauses:

anaʔ	lʊh	tuwɨ	ane	tipʊɪn	tiyaŋ	ʤalan	ʤalan	aʤaʔ	ʧiʧiŋne -
person	*female*	*old*	*that*	*saw*	*I*	*walk*		*with*	*dog-her*

The old woman that I saw took a walk with her dog.

bayine	ane	maɾi	bɨsɨ	mɨʤalan	into
baby-the	*that*	*just*	*can*	*walk*	*this*
ŋɨliŋ	dɔgɛn	kɨranɨ	iyɨ	nyakitaŋ	basaŋ
cries	*always*	*because*	*s/he*	*sickened*	*stomach*

The baby that can just walk always cries because he has a stomach pain.

muɾd	ane	bɨlɔg	into	tɨn	mɨnɛʔ	kɨlas
student	*that*	*stupid*	*this*	*not*	*pass*	*class*

this student that's stupid didn't pass the class

pisagan	˙tiyaŋe	ane	ŋɔyɔŋ	pɨdidinɨ	ʧaɾi	anaʔ	budʊh
neigbor-of	*mine*	*that*	*lives*	*alone*	*looks*	*person*	*crazy*

My neighbor that lives alone looks like a crazy person.

b) conditionals:

lamɨn	baʤune	sampʊn	kɨlit,	baaŋ	adine	dɔgɛn
if	*clothes-the*	*already*	*tight*	*give*	*brother-the*	*(just)*

if the clothes are already tight, just give them to the brother

lamɨn	subɨ	sanʤɨ,	mulɪh	nyin
if	*already*	*evening*	*come home*	*please*

If it's already evening, come home.

lamɨn	umbah	tiyan,	baʤune	niki	bisɨ	dadi	kɨdas
if	*wash*	*I,*	*clothes-the*	*here*	*can*	*become*	*clean*

if I wash, these clothes can be clean

lamɨn	sampʊn	tɨkid	ditu,	kɪnmɪn	nyɨn	tiyaŋ	surat
if	*already*	*arrive*	*there,*	*send*	*please*	*me*	*letter*

if you already arrive there, send me a letter

lamɨn	yɛhe	di	pantʃine	sampʊn	mɨluwab,
if	*water-the*	*in*	*pan-the*	*already*	*boiling,*

pulaŋ	nyɨn	ʤukute
put in	*please*	*vegetables-the*

if the water in the pan is already boiling, put in the vegetables

APPENDIX 5: Focus

example 1:

possible answer:

iyɨ	sidiŋ	ŋae	urutan
s/he	*progressive*	*make*	*sausage*

S/he is making sausage

questions:

a)	iyɨ	sidiŋ	ŋuʤaŋ?	
	s/he	*progressive*	*do*	

What's s/he doing? should elicit a non-focussed reply.

b)	iyɨ	sidiŋ	ŋae	ʤaʤɨ?
	s/he	*progressive*	*make*	*cookies*

Is s/he making cookies? should elicit a reply with narrow focus on the object.

c)	iyɨ	sidiŋ	ŋaʤɨŋ	urutan?
	s/he	*progressive*	*eat*	*sausage*

Is s/he eating sauage? should elicit a reply with narrow focus on the verb.

example 2:

possible answer:

iyɨ	lakar	ŋɨʤet	umah
s/he	*future*	*paint*	*house*

S/he will paint the house.

questions:

a)	iyɨ	sidiŋ	ŋuʤaŋ?	
	s/he	*progressive*	*do*	

What's s/he doing? should elicit a non-focussed reply.

b)	ragane	lakar	ŋɨtʃet	umah?
	you	*future*	*paint*	*house*

Will you paint the house? should elicit a reply with narrow focus on the subject

c)	iyɨ	lakar	mɪnahɪn	umah
	s/he	*future*	*fix*	*house*

Will s/he fix the house? should elicit a reply with narrow focus on the verb.

d)	iyɨ	sampʊn	ŋɨtʃet	umah?
	s/he	*already*	*paint*	*house*

Did s/he already paint the house? should elicit a reply with narrow focus on tense.

e)	iyɨ	lakar	ŋɨtʃɛt	pagɨhan?
	s/he	*will*	*paint*	*fence*

Will s/he paint the fence? should elicit a reply with narrow focus on the object.

REFERENCES

Barber, C.C. (1977) A Grammar of the Balinese Language. Aberdeen University Library, Occasional Publications no. 3. University of Aberdeen.

Bolinger, Dwight. (1978). Intonation Across Languages. in Universals of Human Language. Joseph H. Greenberg, ed. vol.2. Stanford University Press: Stanford, CA.

Comrie, B. (1985) Causative Verb Formation and Other Verb-Deriving Morphology. chapter 6 in Language Typology and Syntactic Description, vol. 3. Grammatical Categories and the Lexicon. Timothy Shopen, ed. Cambridge: Cambridge University Press.

Geertz, C. (1972) Linguistic Etiquette. chapter 11 in Sociolinguistics. J.B. Pride and Janet Holmes, eds. Penguin Modern Linguistics Readings.

Horne, E.C. (1961) Beginning Javanese. Yale University Press: New Haven.

Jun, S-A. (1993) The Phonetics and Phonology of Korean Prosody. PhD Dissertation. The Ohio State University.

Jun, S-A. and Fougeron, C. (1995) The Accentual Phrase and the Prosodic Structure of French. in Proceedings of the XIIIth International Congress of Phonetic Sciences. Kjell Elenius and Peter Branderud, eds. vol. 2, pp. 722-725. Stockholm, Sweden.

Jun, S-A. and Oh, M. (1994) A Prosodic Analysis of Three Sentence Types with "Wh" Words in Korean. in Proceedings of the International Conference on Spoken Language Processing. The Acoustical Society of Japan. pp. 323-326.

Laksman, Myrna. (1994) Location of Stress in Indonesian Words and Sentences. in Semaian 9. Experimental Studies in Indonesian Prosody. Cecilia Odé and Vincent J. van Heuven, eds. Vakgroep Talen en Culturen van Zuidoost-Asië en Oceanië: Rijksuniversiteit te Leiden.

Maekawa, K. (1991) Perception of Intonational Characteristics of Wh- and non-Wh-Questions in Tokyo Japanese. in Proceedings of the International Congress of Phonetic Sciences.

Odé, C. (1994) On the Perception of Prominence in Indonesian. in Semaian 9. Experimental Studies in Indonesian Prosody. Cecilia Odé and Vincent J. van Heuven, eds. Vakgroep Talen en Culturen van Zuidoost-Asië en Oceanië: Rijksuniversiteit te Leiden.

Schachter, P. (1985). Parts-of-Speech Systems. chapter 1 in Language Typology and Syntactic Description, vol. 1. Clause Structure. Timothy Shopen, ed. Cambridge: Cambridge University Press.

Selkirk, E.O. (1986) On Derived Domains in Sentence Phonology. Phonology Yearbook 3. pp. 371-405.

Shadeg, N. (1977) A Basic Balinese Vocabulary. Dharma Bhakti, Denpasar.

Stevens, A. M. (1965) Language Levels in Madurese. Language 41:2.

Ward, J.H. (1973) Phonology, Morphophonemics, and the Dimensions of Variation in Spoken Balinese. PhD Dissertation, Cornell University.

OSU Working Papers in Linguistics 50 (101-113)

The auditory/perceptual basis for speech segmentation [*]

Keith Johnson
kjohnson@ling.ohio-state.edu

Abstract: Language is temporal in two ways. Words and sentences occur in time, each utterance having a beginning and end. But, also the learner's experience of language occurs over time, the items that are crucial for defining linguistic structure are experienced over the course of years. These two observations are addressed in an exemplar model of phonological learning and word recognition. Major features of the model are described and its operation is illustrated in two simulations.

Introduction

Language unfolds slowly over time. Of course, sentences and words are temporal events to be segmented and analyzed. But in addition to this local temporal structure, the experience of language as a whole occurs over time. Only rarely is an explicit phonological contrast demonstrated to the child. Most elements of language structure if they are to be learned must be extracted from memory. That is to say, the contrasts and similarities, the crucial comparisons from which linguistic structure emerges, are based on remembered instances of linguistic objects.

Consider the role that the linguist's 3x5 cards, notebook, or relational database plays in producing a linguistic analysis. The cards are used to write transcriptions of words, which are drawn from work with consultants who teach the linguist how to say things in a given language. These records are the starting point for linguistic analysis. They are culled and compared, stacked according to similarities. Words with very similar pronunciations but very different meanings reveal phonemes, the minimally contrastive sounds in the language, and words with slightly divergent pronunciations but similar meanings form paradigms, revealing patterns of inflection or word derivation.

The point is that, for both child and linguist, linguistic structure - the analysis of language into its combinable elements - crucially relies on a pre-analytic store of linguistic items in memory.

This is one of the considerations which has led me to explore a class of models called instance-based or exemplar models of linguistic memory and speech recognition. Before going on to describe some simulations of the process by which linguistic structure emerges from specific instances in memory I will briefly outline some further considerations which point to an instance-based model of speech recognition.

[*] This paper was presented as a part of the OSU Seminar on Human Communication sponsored by the Cognitive Science Program, Department of Linguistics, Department of Psychology, and Department of Speech and Hearing Science, on March 31, 1997. I am grateful to the organizers of the seminar, Osamu Fujimura, Rob Fox, and Neal Johnson for inviting me to speak. For helpful discussions on various aspects of this work I would like to thank Anne Cutler, Mary Beckman, Mark Pitt, Rob Fox, and the members of the OSU Linguistics Laboratory. Thanks also to Kim Ainsworth-Darnell for helpful comments on an earlier version of the paper.

Exemplars in speech processing

A traditional concern in the theory of speech perception is that phonemes vary quite considerably across talkers and contexts. That is, the acoustic cues for phonemes lack invariance, and consequently pose a difficult problem for theories of speech recognition (and for automatic speech recognition systems).

It turns out that variation across talkers can be reduced by normalization schemes (see figure 1). For example, Potter & Steinburg (1952) noted that the ratios of vowel formant frequencies show much less variation across talkers than do their absolute values. Observations such as this have led researchers (Bladon, Henton & Pickering, 1984; Miller, 1989; Syrdal & Gopal, 1986; Sussman, 1995; Traunmüller, 1981) to assume that linguistic categories are recognized by reference to 'higher order invariants' like formant ratios. (Gibson, 1966, was especially influenced by this argument.)

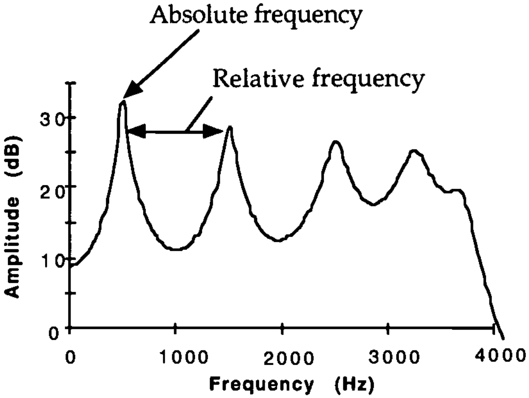


Figure 1. Formant ratios as higher order invariants. Distances between formants show less between-talker variability than to absolute formant frequencies.

So, the basic scheme for recognition in this view has two stages, (1) normalization then (2) comparison with category prototypes. Simply put, this approach doesn't work. When you implement it you get recognition performance that doesn't come close to human performance, and then you have to build separate mechanisms to recognize speakers, dialects, styles, and so on. Miller (1989, see figure 2)) showed that the shapes of the category regions in a derived 'higher order' perceptual space are quite irregular, and adequate performance is best achieved by demarking category regions by reference to exemplars. The vowel regions that Miller (1989) presented show a multimodal structure even in the 'higher order' space.

This "normalize and compare" scheme doesn't work because Potter & Steinburg's observation is only approximately true. Talker differences are only partly eliminated in higher-order invariants and the remaining differences are enough to disrupt recognition by reference to prototypes. This is true even for very constrained laboratory speech such as in the Peterson & Barney (1952) database. If we consider even small variations in speaking styles (isolated words versus words in carrier phrases) we see further overlap and multimodal distributions.

Another consideration is the fact that recent research has found that prior exposure to an utterance facilitates later recognition. For example, Goldinger (1997) found evidence for the retention of word exemplars in tests of implicit memory. If the identity of the talker was the same across repetitions of a word in successive blocks of word recognition trials

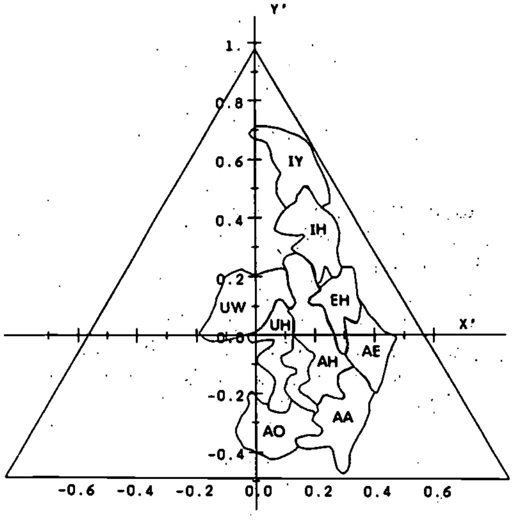


Figure 2. Miller's (1989) auditory/perceptual space. Category regions in a 'higher order invariant' representation are irregularly shaped; category boundaries are determined by exemplars near the boundary.

(even if the blocks occurred one week apart) listeners were able to recognize the word more accurately than if the repeated word was spoken by a different talker. A gain in word recognition accuracy across repetitions in the two blocks of trials occurred in all nine conditions in Goldinger's study, varying over delay intervals and number of talkers. These results (and others like them) show that low-level acoustic details of word presentations are retained and have an effect on later processing.

Finally, neurophysiological studies of memory show that single events alter synaptic strengths, and even the number of synapses, in the hypocampus. I don't want to make too much of this other than to note that I take these studies to indicate that long-term changes in neurological organization though perhaps small may result from single events. These findings lend a bit of plausibility to an exemplar-based model of speech recognition, however indirectly relevant they may be in other ways.

Whole-word exemplars

The Goldinger (1997) study and and other work along the same lines suggest that remembered instances of speech are acoustically detailed as would also be expected on psychophysical grounds. Additional evidence suggests that not only are speech exemplars acoustically detailed, but at least during language acquisition they are also unanalyzed whole words.

Bregman (1990) outlines several principles of primitive auditory scene analysis, which break an auditory array into objects - the fan of the projector, the cough at the back of the room, the utterances of the talker, and so on. In this view, these auditory objects have beginnings and ends, but no internal structure. Assuming that speech recognition begins with auditory scene analysis, we then would have to say that speech recognition starts with unanalyzed wholes.

The 'holistic' stage in language acquisition shows that children at first learn words without internal structure. One reason to believe that this is the case is that during acquisition there is a period of rapid vocabulary growth (often called an 'explosion') which suggests that the child has learned to analyze auditory objects into recombinable articulatory primitives. This sudden change in behavior is evidence that at an earlier stage language was not so organized.

A related observation is that phonological inventories and alternations are position specific. For example, the inventory of contrastive vowels in English depends on context; in my dialect there are 10 contrasting vowel qualities in [hVd] context but only 6 in [hVr]. Also, in many languages consonant voicing contrasts are 'neutralized' in coda position, and in all languages the acoustic cues for consonants differ between onset and coda. These facts are relevant to the view that speech exemplars are unanalyzed wholes because they show that similarity and contrast depend upon temporal location within utterances.

Assumptions

In summary the work described in this paper starts from three assumptions. First, speech is recognized by reference to stored instances (exemplars). Second, these exemplars have no internal structure, rather they are unanalyzed auditory representations. And third, they are word-sized chunks, as a result of primitive auditory scene analysis where isolated word productions form the basis for word recognition in running speech.

The reader may prefer to think of these assumptions as relevant for the development of linguistic representations during language acquisition, but the evidence suggests that adults also use exemplars in word recognition. So keep in mind as we discuss some simulations that these properties may be active in the mature recognition system.

How is speech 'analyzed' into segments?

Given these assumptions, in particular that remembered instances of speech are stored as unanalyzed wholes, how can speech be analyzed into segments?

One answer is that it isn't, that the segmentation of speech is a figment of the imagination fired by orthography. Some reasons to believe that this stance is incorrect have already been mentioned. The 'lexical explosion' argument, for example, is evidence for both preanalytic representation and of segmentation. Three additional observations suggest that segmentation is not merely an invention.

Listeners and talkers experience the speech stream as a sequence of separate words, any one of which can be repeated or replaced. Though a model that assumes word-sized exemplars may readily handle such segmentation (see Johnson, 1997), it should be noted that primitive auditory scene analysis does not. To achieve word-level segmentation in running speech we must posit a system in which word-sized exemplars support a cognitive scene analysis in which the recognition system segments the speech stream into words.

We can also note briefly that writing systems generally reflect analyses of speech into recombinable units such as segments or syllables, and their very existence suggests the psychological reality of sublexical units at some point in history, though not necessarily the use of these units in on-line speech processing. Also, segmental speech errors suggest that segmental organization is used in speech production.

The remainder of this paper describes an exemplar-based model of auditory word recognition - focusing particularly on behavior of the model which is related to segmentation. These simulations explore the degree to which linguistic structure may be an emergent property of recognition based on remembered auditory representations of speech.

The model

The basic operation of an exemplar model (Nosofsky, 1986) is to categorize perceptual objects by evaluating the similarity between the item to be categorized and a set of stored category exemplars. Within-category variation is explicitly represented in the set of exemplars (which substantially out-number the categories). Similarity between exemplars and the unknown is an exponential function of auditory distance (1) where d_{ij} is the Euclidian distance between exemplar *j* and the unknown object *i*. and κ is a sensitivity parameter. Word activation is the weighted sum of similarity (2) where W_{jc} is the connection weight between exemplar *j* and word *c*..

$$sim_{ij} = exp(-\kappa d_{ij}) \qquad (1)$$

$$act_c = \Sigma sim_{ij} * W_{jc} \qquad (2)$$

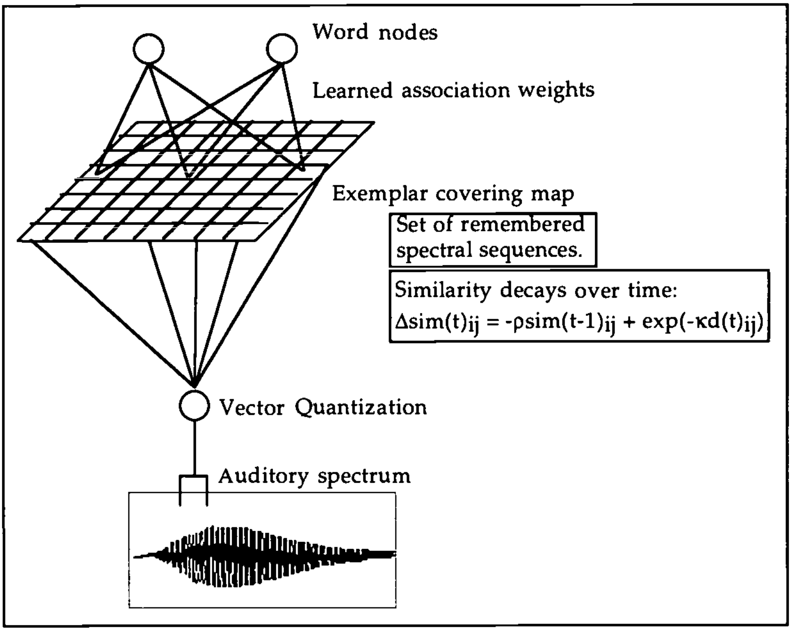

Word nodes
Learned association weights
Exemplar covering map
Set of remembered spectral sequences.
Similarity decays over time:
$\Delta sim(t)_{ij} = -\rho sim(t-1)_{ij} + exp(-\kappa d(t)_{ij})$
Vector Quantization
Auditory spectrum

Figure 3. An exemplar model of auditory word recogntion. Processing proceeds from the bottom to the top of the graph. Each 23 ms frame of speech is processed by an auditory model, vector quantized, and compared with the set of remembered sequences in the exemplar covering map. Word node activation is the product of similarity to the covering map location and the learned associations between that location and the word node.

The model has three stages of processing (see figure 3). The first stage converts the speech wave form into a sequence of auditory spectra. I use a very simple psychoacoustic critical-band filtering routine (Johnson, 1990), with a frame rate of 43 Hz. The auditory spectra are then coded as most similar to one of a set of stored spectra. This vector quantization stage is not strictly necessary for the simulations that I discuss here and introduces some noise, but is necessary for systems that store a large number of exemplars because with vector quantizing each spectrum in an exemplar can be stored as a single integer rather than as a vector of real numbers. The vector-quantizing stage uses adaptive resonance theory (Carpenter & Grossberg, 1989). The in-coming spectrum is compared with the spectra stored in the vector quantizing codebook and if it is not similar to any of them it is added to the codebook, given a code number, and that number is returned. If it is similar to one of the stored spectra, that spectrum's code number is returned and the stored spectrum is shifted slightly to be more similar to the in-coming spectrum. The degree of noise introduced by vector quantizing is thus determined by a 'vigilance' parameter which determines when spectral templates will be added to the VQ codebook. This approach is also used in the third stage during training but not during test to build an exemplar covering map.

In the third stage the sequence of auditory spectra is compared with the sequences stored in an exemplar covering map (Kruschke, 1992). As with vector quantizing, if the in-coming sequence is unlike any existing exemplar it is added to the map. Weights connecting locations in the exemplar covering map and word nodes are learned during training by counting the number of times that the covering map location is an instance of the word. The weight connecting the closest exemplar in the covering map and the 'correct' word is incremented by one for each training token. It should be noted that though the architecture shown in figure 3 is similar to Kruschke's (1992) ALCOVE model, the use of exemplars is more similar to Nosofsky's (1986) GCM.

The model assumes that similarity is evaluated one time-frame at a time starting at the onset of the auditory objects being compared, and that activation decays over time as a function of a decay parameter ρ:

$$sim(t)_{ij} = -\rho sim(t-1)_{ij} + exp(-\kappa d(t)_{ij}) \qquad (3)$$

Simulation of the recognition of 'cap'.

This model was trained to recognize eight words (Table 1) spoken in list-reading style by a single male speaker. The utterances were recorded directly to computer disk at a 22 kHz sampling rate, using 16 bit samples. The words were chosen to illustrate segmental contrasts in CVC words, and a case of a 'phantom' word. 'Catalog' and 'battle-log' may be confusing for a word recognition system because right context '-alog', '-le-log' distinguishes the short words 'cat' and 'bat' from the longer words 'catalog' and 'battle-log'. The system being tested in this simulation is not time invariant, but rather assumes that the beginnings and endings of the words are known. Nonetheless, we will see some interesting behavior in the segmentation of the longer words.

Table 1. Words used in the first simulations.

bap	cap
bat	cat
battle-log	catalog
beet	keep

The model was trained on the first 10 of 13 repetitions of each word. During training the codebook and exemplar covering map were constructed and weights between exemplars and words were established in one pass through the first 10 repetitions of the words. This very simple training algorithm led to 96% correct recognition of the remaining three repetitions of the words.

Figure 4 shows word activations as a function of time (which is given in frame number) during the presentation of 'cap' to the model. A spectrogram of the instance of 'cap' being recognized is shown in approximate alignment with the frames. At frame 1 all of the words starting with [k] are more activated than are the [b] words. At frame 3 activation of 'keep' drops out, giving us a set of activated words that start [kæ]. At frame 5 all of the words with [æ] as the first vowel show increasing activation. At frame 10 activation for 'catalog' drops off (as did 'battle-log' at frame 7), perhaps because of vowel duration mismatches. At frames 13-16 the words that end in [p] show increasing activation, though this increase is only slight for the correct answer 'cap'.

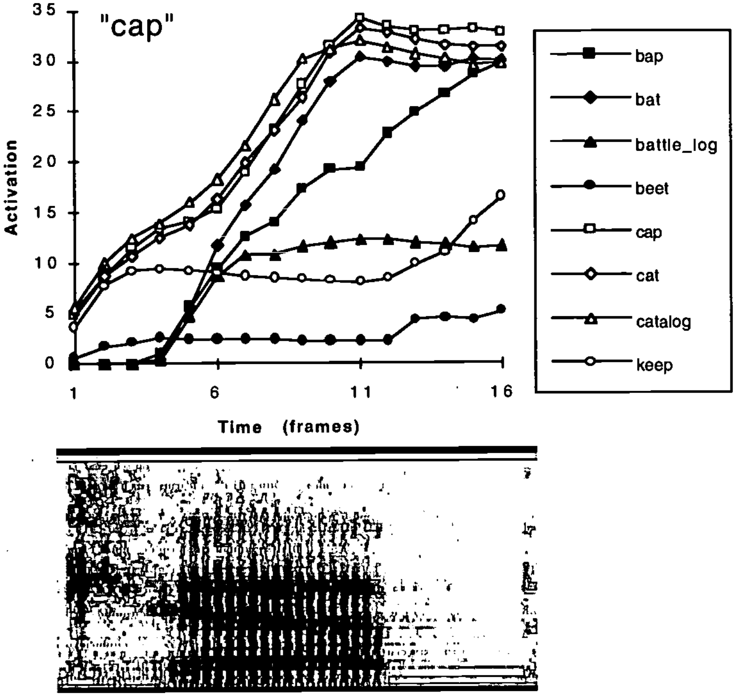


Figure 4. Spectrogram of the word 'cap' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top).

One interesting aspect of this simulation (and of the one to follow) is that right context cues seem to exist in the onset syllable 'catalog'. Because pronunciations of a sequence like [kæt] differ phonetically in monosyllabic and disyllabic words, a model such as the one discussed here which is sensitive to acoustic detail begins detecting the difference between them even during the initial CVC sequence. One argument supporting lexical competition in models like TRACE (Elman & McClelland, 1986; McClelland & Elman, 1986) and Shortlist (Norris, 1994) is that the initial sequence of phonemes in 'catalog' is no different than the sequence 'cat'. This simulation suggests though that word pairs such as this which seem to require right context for disambiguation are not completely ambiguous without right context. Hence, some of the work done by lexical competition is accomplished by attention to phonetic detail in the exemplar model.

These patterns of activation also show three emergent segments. /k/ emerges in the first frame as the subset of lexical items which begin with [k], /æ/ emerges at frame 5 as the set of words having first vowel [æ], and /p/ emerges near the end of the word as the words with [p] codas show increasing activation. This analysis suggests that phonemes are defined in terms of subsets in the set of exemplars which have time-aligned similarities in their auditory/perceptual representations.

This account of the emergence of segments from unanalyzed exemplars has some interesting properties. The emergent segments are based on auditory similarity and are position-specific. They are not, however, Wicklephones (Wicklegren, 1969). That is, we see here evidence that all words with [æ] show increasing activation during the vowel, not just those that have the same consonantal context. (We will see below a case in which a context sensitive allophone emerges.)

Though I am interpreting the pattern of activation in 'cap' as having emergent segments, there are no segmental representations in memory being activated in response to the signal. That is, no recombinable units exist in the memory representations of the words. This is because I have stored exemplars as containing auditory descriptions only, as formalized in (4) where E_O is a set of exemplars defined by auditory properties A. If we assume that the child's own productions are exemplars that contain both auditory and motor descriptions, as formalized in (5) where E_S is a set of exemplars defined by both auditory properties A and motor commands M, we can speculate that the sequence of activated lexical subsets that we have just seen gives rise to the activation of a sequence of articulatory gestures.

$E_O = \langle A \rangle$ exemplars produced by others (4)

$E_S = \langle A,M \rangle$ exemplars produced by self (5)

Simulations of the recognition of 'catalog' and 'battle-log'

We turn now to simulations of the recognition of 'catalog' and 'battle-log' using the same vocabulary and trained model that were just described.

Figure 5 shows word node activation levels over time in response to an instance of the word 'catalog'. Many of the segmental phenomena that we saw in the 'cap' example are evident here as well. For example, as before in the first frame all of the [k] initial words show increased activation in response to the word 'catalog'. Also in frame 6 all of the [æ] words show increasing activation.

But in addition to these segmental phenomena we see at the end of the word (frame 16 and after) that both 'catalog' and 'battle-log' show increasing activation at about the same rate over time. In this case the unit of linguistic structure which is being defined by a lexical subset is a syllable. A hierarchical structure of syllables and segments emerges from the activity of the model.

This is apparent also in the word activations in response to the word 'battle-log' which are shown in figure 6. Some segmental phenomena are seen during the first syllable while over the course of the second syllable both 'catalog' and 'battle-log' show increasing activation.

Figure 6 also shows the context sensitive allophonic response that was mentioned earlier. In the first frame only the three [bæ] words are activated. The word 'beet' remains virtually unactivated during the entire course of the word 'battle-log', despite the fact that they both start with 'b'. This is a topic for future investigation, but this simulation does suggest that there may be circumstances in which the subsets of activated lexical items generated by the model define allophones rather than phonemes or syllables.

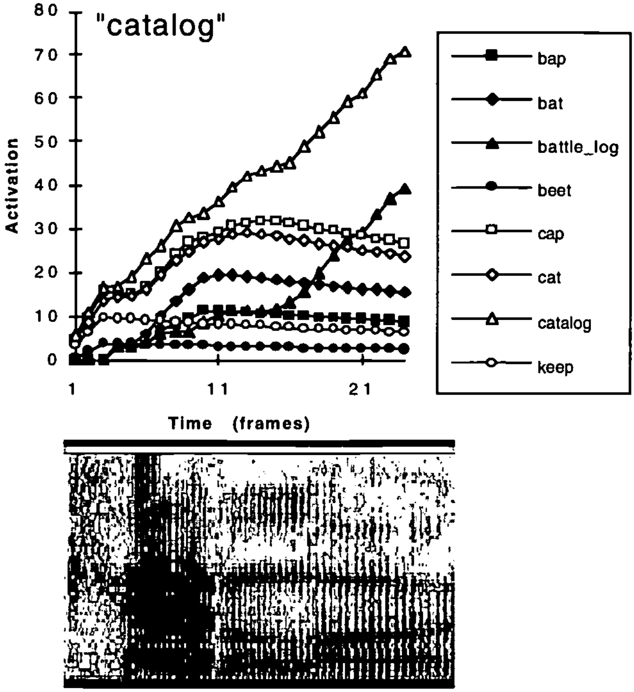


Figure 5. Spectrogram of the word 'catalog' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top).

Simulating the metrical segmentation strategy

The final simulations use a different database of training tokens and then replicates Cutler and Norris' (1988) finding that it is easier to spot the word 'mint' in a nonword with a strong-weak metrical structure like 'mintuf' ['mɪntəf] than it is in a nonword with a strong-strong metrical structure like 'minteif' ['mɪn,teɪf]. The model was trained on the words listed in Table 2 as before and then tested on the nonwords. Each word in table 2 was repeated eight times in isolation by a single male talker and recorded directly to computer disk with 22 kHz, 16 bit sampling.

Table 2. Words used in the second simulations.

training words:		test nonwords:	
mint	men		
minty	rented	mintuf	minteif
mints	minted		
retain	maintain		

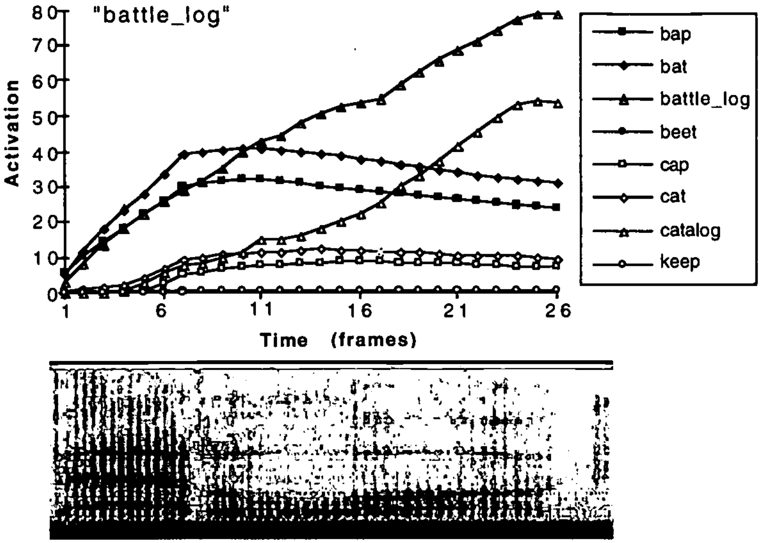


Figure 6. Spectrogram of the word 'battle-log' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top).

Figure 7 shows the activations of the eight words in the lexicon in response to the non-word 'mintuf', which is shown in the spectrogram. In frame 1, activations of the words that start with [m] are greater than those that start with [r]. The most highly activated word in the eight word lexicon is 'minted', and during the first syllable 'mint' and 'mints' show relatively high activation (near a value of 25). By the end of the word there is a cluster of relatively activated words having activation somewhat less than 'minted'; these were 'mint', 'mints','minty', and 'maintain'.

Now compare this with the pattern of activations prompted by 'minteif' (Figure 8). Some segmental phenomena are apparent in this simulation. As before, in frame 1 words starting with [m] are more activated that words starting with [r]. Also, as in figure 7 the final fricative in 'mintuf' seems to have partially overlapped with the final fricative in 'mints' indicating that the model is sensitive to mid-class phonetic similarity (Dalby, et al., 1986).

The most highly activated word in the lexicon was 'maintain' which like this production of 'minteif' has two metrically strong syllables. The other words which show fairly high activation in response to 'minteif' are the two syllable words in the lexicon which have a strong first syllable. Interestingly, 'retain' which was pronounced by this speaker with a weak first syllable only showed increasing activation during the second syllable of 'minteif'.

Finally, note that the activation of 'mint' peaks at about 15. Given that the activation of 'mint' reached 25 in response to 'mintuf' we would predict that it would be easier for the model to spot 'mint' in 'mintuf' just as it was for Cutler and Norris' subjects.

This simulation, in addition to modeling Cutler and Norris' result without explicitly segmenting the speech stream into metrical feet, shows that like segments and syllables, metrical units may emerge as sets of activated lexical items in an exemplar-based recognition model.

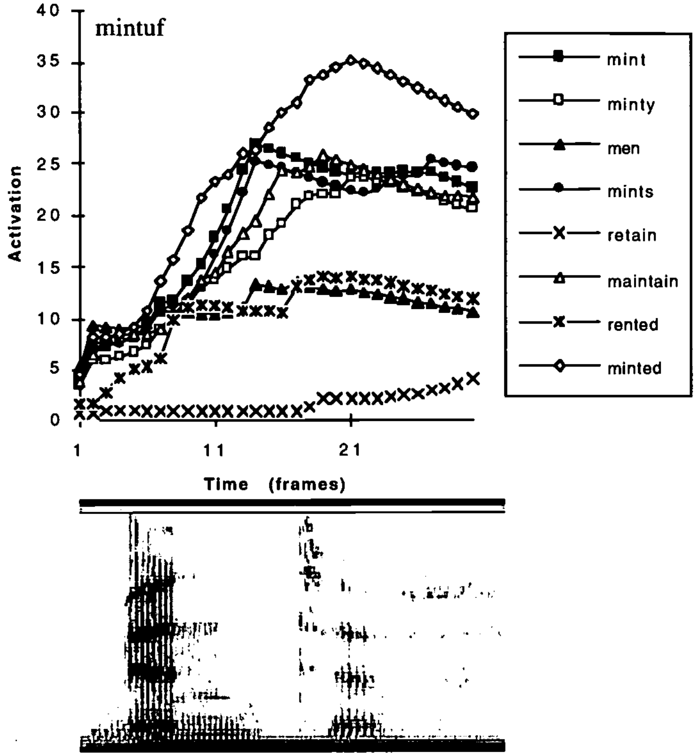


Figure 7. Spectrogram of the nonword 'mintuf' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top)..

Conclusion

One way to describe these results is say that they describe a developmental process by which a child learning language might build representations of abstract linguistic units like segments, syllables, or metrical feet. Indeed, I was inspired in this line of research by a talk by Jan Edwards on phonological disorders in language acquisition.

However, if you accept any of the arguments supporting the view that adult speech recognition is an exemplar-based process, then we can raise the interesting possibility that abstract phonological structure is a fleeting phenomenon - emerging and disappearing as words are recognized.

This may explain what we mean when we say that the speaker/hearer has implicit or unconscious knowledge of phonological structure. Abstract phonological structure in this view is never explicitly stored or detected, though the subsets of lexical items which define these abstract entities, for both the language user and the linguist, are implicitly linked through their auditory/perceptual similarities.

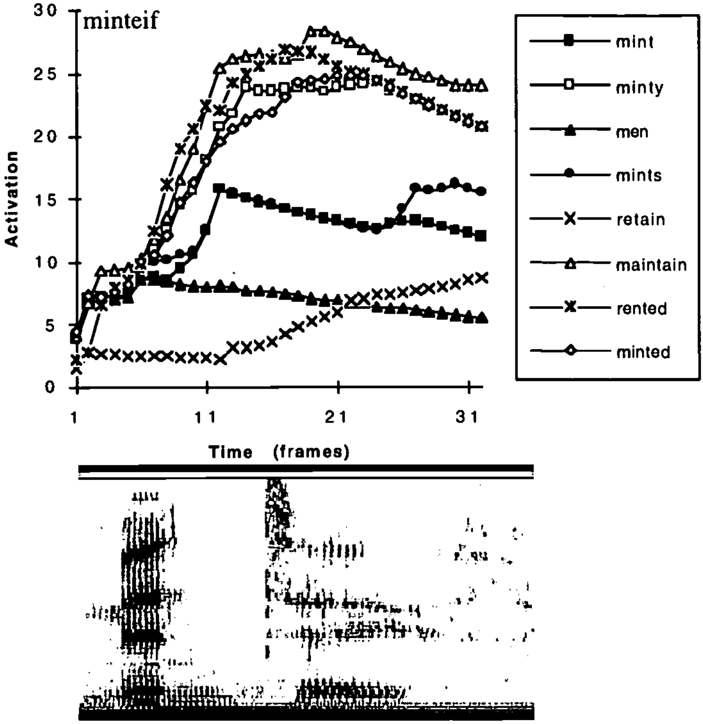


Figure 8. Spectrogram of the nonword 'minteif' (bottom) with word activations produced by the model approximately time-aligned with the spectrogram (top).

References

Bladon, A., Henton, C. & Pickering, J.B. (1984) Towards an auditory theory of speaker normalization. *Language Commun.*, 4, 59-69.

Bregman, A. (1990) *Auditory Scene Analysis.* Cambridge: MIT Press.

Carpenter, G.A. & Grossberg, S. (1989) Search mechanisms for adaptive resonance theory (ART) architectures. *International Joint Conference on Neural Networks*, Washington, DC, June, 18-22, 1989 (Vol I, pp. 201-205). Piscataway, NJ: IEEE.

Cutler, A. & Norris, D. (1988) The role of strong syllables in segmentation for lexical access. *J. Exp. Psych.: Hum. Perc. & Perf.*, 14, 113-121.

Dalby, J., Laver, J. & Hiller, S.M. (1986) Mid-class phonetic analysis for a continuous speech recognition system. *Proceedings of the Institute of Acoustics*, 8, 347-354.

Elman, J. & McClelland, J. (1986) Exploring lawful variability in the speech waveform. In S. Perkell & D.H. Klatt (Eds.) *Invariance and Variability in Speech Processing* (pp.

360-385). Hillsdale, NJ: Erlbaum.

Gibson, J.J. (1966) *The Senses Considered as Perceptual Systems*. Boston: Houghton-Mifflin.

Goldinger, S.D. (1997) Words and voices: Perception and production in an episodic lexicon. In Johnson, K. & Mullennix, J.W. (Eds.) *Talker Variability in Speech Processing* (pp. 33-66). NY: Academic Press.

Johnson, K. (1990) Contrast and normalization in vowel perception. *J. Phon.* **18**, 229-254.

Johnson, K. (1997) Speech perception without speaker normalization. In Johnson, K. & Mullennix, J.W. (Eds.) *Talker Variability in Speech Processing* (pp. 145-166). NY: Academic Press.

Kruschke, J. (1992) ALCOVE: An exemplar-based connectionist model of category learning. *Psych. Rev.*, **99**, 22-44.

McClelland, J. & Elman, J. (1986) The TRACE model of speech perception. *Cog. Psych.*, **18**, 1-86.

Miller, J. (1989) Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am.*, **85**, 2114-2134.

Norris, D. (1994) Shortlist: a connectionist model of continuous speech recognition. *Cognition*, **52**, 189-234.

Nosofsky, R.M. (1986) Attention, similarity, and the identification-categorization relationship. *J. Exp. Psych.: Gen.*, **115**, 39-57.

Peterson, G.E. & Barney, H.L. (1952) Control methods used in a study of the identification of vowels. *J. Acoust. Soc. Am.*, 24, 175-184.

Potter, R. & Steinburg, J. (1950) Toward the specification of speech. *J. Acoust. Soc. Am.*, **22**, 807-820.

Sussman, H.; Fruchter, D. & Cable, A. (1995) Locus equations derived from compensatory articulation. *J. Acoust. Soc. Am.*, **97**, 3112-3124.

Syrdal, A. & Gopal, H. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. Acoust. Soc. Am.*, **79**, 1086-1100.

Traunmüller, H. (1981) Perceptual dimension of openness in vowels. *J. Acoust. Soc. Am.*, **69**, 1465-1475.

Wickelgren, W.A. (1969) Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psych. Rev.*, **76**, 1-15.

Production and perception of individual speaking styles [*]

Keith Johnson & Mary E. Beckman
kjohnson@ling.ohio-state.edu
mbeckman@ling.ohio-state.edu

Abstract: As explanation of between-speaker differences in speech production moves beyond sex- and age-related differences in physiology, discussion has focused on individual vocal tract morphology. While it is interesting to relate, say, variable recruitment of the jaw to extent of palate doming, there is a substantial residue of arbitrary differences that constitute the speaker's "style". Style differences observed across a well-defined social group indicate group membership. Other style differences are idiosyncratic "habits" of articulation, individual solutions to the many-to-many mapping between motoric and acoustic representations and to the many different attentional trading relationships that can exploit the typical patterns of redundant variation in independent acoustic correlates of any minimal contrast. Perceptual studies of social style differences suggest that perceptibility depends upon the task and upon the hearer's own group membership. The few studies of idiosyncratic differences suggest that speakers perceive each others' productions in terms of their own habits. Thus, perceptual compensation for speaker differences must go beyond mere vocal tract normalization. A promising route for describing how listeners compensate for the arbitrary variation of style is an instance-based (or exemplar) model of speech perception in which the distribution of exemplars is heavily weighted by instances of the speaker's own productions.

Introduction

Until very recently, most discussion of between-speaker differences has been couched in the framework of "speaker normalization". In this framework, gestures are equated with the dimensions of invariant linguistic contrast between phonemes ("distinctive features"), and between-speaker variability is treated as an artifact of the transmission line — a kind of noise which needs to be filtered out of the signal in order to get at the meaningful category variation.

This paper illustrates several ways in which gestures for the same phoneme category can differ meaningfully across speakers, and then discusses the implications for our models of the listener. If listeners can categorize speakers, then the problem is not merely one of normalizing over speakers to perceive phonemes, but a more general problem of how to extract categories in one dimension of classification in the face of meaningful variation in another dimension of classification. We propose a model of how listeners might process utterances for all of the linguistically relevant categories that the signal encodes.

[*] Talk presented at the 3rd joint meeting of the Acoustical Society of America and the Acoustical Society of Japan, Honolulu, 2-6 December, 1996. Thanks to Mariapaola D'Imperio for helpful comments on an earlier version.

Meaningful Variation

The earliest work on between-speaker differences, of course, categorized speakers entirely in terms of age- and sex-related changes in vocal tract morphology (see Peterson & Barney, 1952 and use of these data in testing nearly all subsequent proposals of vowel normalization algorithms). In particular, we know that adult male talkers tend to have lower formants and lower fundamental frequencies than adult females do, because of hormonal changes at puberty that lead both to a descent of the larynx that elongates the vocal tract and a simultaneous change in the morphology of the thyroid cartilage that elongates the vocal folds. Such observations prompted algorithms for normalizing formant values by fundamental frequencies, and the like, to find the invariant underlying gesture (Nearey, 1978; Miller, 1989).

Articulatory studies, however, suggest that there are real between-speaker differences in gesture. For example, figure 1 shows x-ray traces of jaw and tongue surface at vowel mid-point in front vowels produced by two adult male speakers of American English. The speaker on the left shows a large variation in jaw height that is systematically related to the contrasts between the two high and two mid vowels and between the mid and low vowels. The speaker on the right shows hardly any variation in jaw height across the five vowels. We speculate that these different gestural strategies may be related to between-speaker differences in palate shape. That is, a more steeply domed palate might be associated with an individual articulatory style that does not recruit the jaw much in tongue raising and lowering gestures for vowels.

In figure 2 we see similar between-speaker variation in the coordinated movement of jaw and tongue in a set of magnetometer studies reported by Harrington and Fletcher (1996). They compared high and low vowels of Australian English produced in accented versus unaccented positions in the intonation contour, and showed that some speakers (such as JMF) lower the jaw more in accented syllables, whereas other speakers (such as LML) have very little variation in jaw position across accented versus unaccented position. Here individual speaker style seems to be associated not with different morphologies,

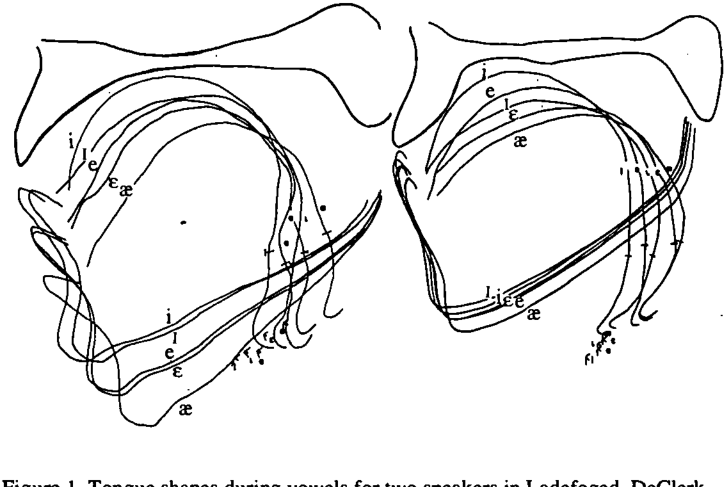


Figure 1. Tongue shapes during vowels for two speakers in Ladefoged, DeClerk, Lindau & Papcun (1972).

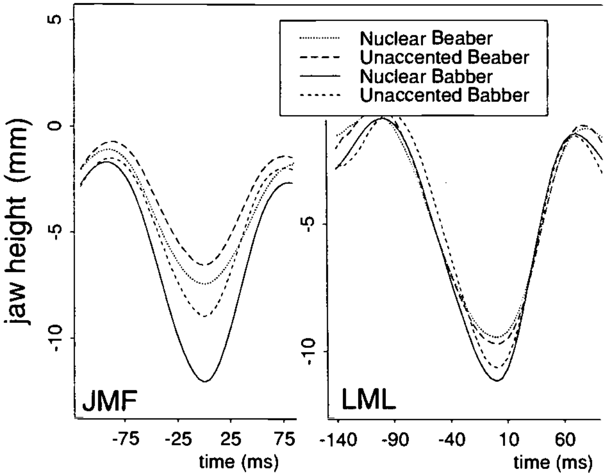


Figure 2. Average jaw trajectories for speakers JMF and LML from Harrington & Fletcher's (1996) study.

but with subtle differences in the prosodic system. JMF consistently makes vowels in accented syllables longer and more peripheral, whereas LML does not use these redundant non-tonal cues to pitch accent placement. See also Edwards, Beckman, and Fletcher (1991) and deJong (1995) for comparable inter-speaker differences in prosodic strategies for American English. Harrington & Fletcher's study also suggested another kind of difference between the two talkers in figure 2. Figure 3 shows traces for the tongue body vertical position and for the first formant in representative tokens of utterances with the high tense vowel [i:] produced by these two talkers. JMF's production has a high tongue body throughout the vowel, and a relatively flat and very low F1, whereas LML's production shows a pronounced diphthongal movement, with a distinct peak in tongue body position late in the vowel and a much higher F1 at the beginning of the vowel. We suspect that this difference is part of a larger pattern of variation in style defined by the continuum of Australian English features. That is, JMF's higher vowel here is typical of so-called "Cultivated Australian", which is closer to British English, whereas LML's lower onset and decidedly diphthongal pattern is closer to the "Broad Australian" end of the continuum.

Figure 4, from Harrington and Cassidy (1994), shows average formant values in a database of Australian English and a comparison plot of typical British English formant values. The diphthongal lowering at the beginning of the tense front vowel in *heed* in Australian English doesn't show up very well, because these averages are from the vowels' midpoint values. But the figure does show another salient feature of Australian English — namely, the raising of the lax front vowels in *hid*, *head*, and *had*. To us, this pattern is very reminiscent of some differences in regional dialects that we've found in our ongoing studies of American English vowel systems.

Figure 5 shows vowel spaces for 13 female talkers from Birmingham, Alabama, and 7 from Los Angeles, California. The Alabama data are from Johnson's unpublished work, and the California data are from Johnson, Flemming, & Wright (1993). As in figure

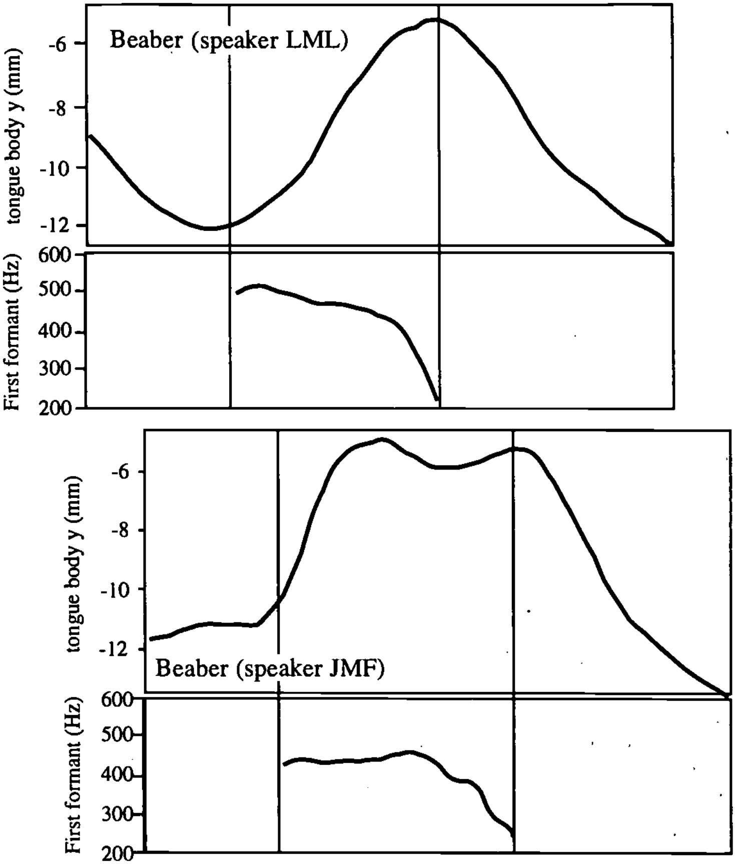


Figure 3. Tongue body height and first formant trajectories of sample tokens of "Beaber" from speakers JMF and LML. Vertical lines mark [i] onset and offset.

4, the formant values that are observed here are taken from vowel midpoints, so the plot does not show that many of the Alabama speakers had a lower onset value for the tense front vowel. However, the figure does show that for the Alabama speakers the lax front vowels in *head* and *had* are raised relative to productions by the Los Angeles speakers. These kinds of speaker-style differences observed across a well-defined social or regional group indicate group membership, and sociolinguistic studies have shown that listeners can be very acutely aware of them, particularly when the differences are associated with differences in social prestige or stigma (see, e.g., Labov, 1966; Trudgill, 1974). A normalization algorithm which treats this kind of between-speaker variability as noise to be factored out of the signal could not be an accurate model of how real listeners extract relevant categories from the signal.

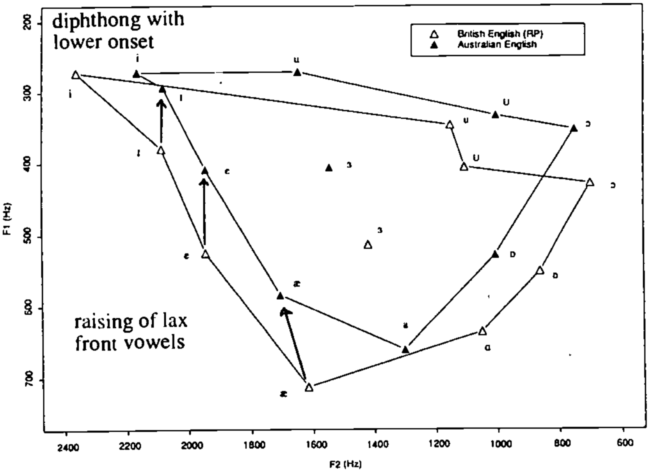


Figure 4. Harrington & Cassidy (1994) acoustic vowel formant measurements comparing Australian versus RP vowel spaces.

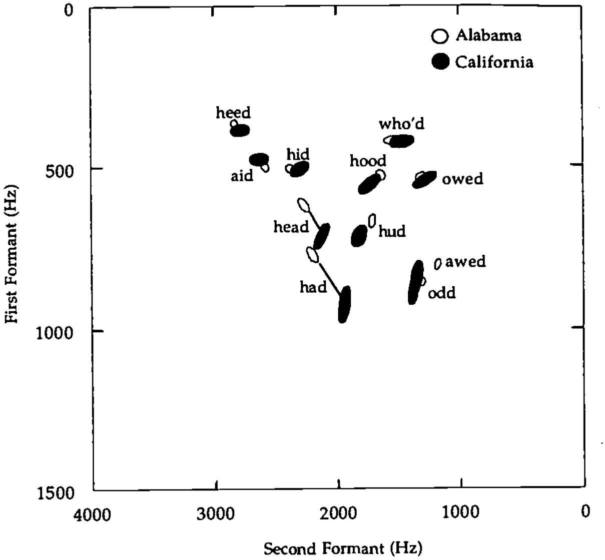


Figure 5. Spectral differences across dialects: Alabama versus Los Angeles vowel spaces. Ellipses show 95% bivariate confidence intervals: filled - Los Angelenos, open - Alabamians.

Using Variability in Speech Perception

We have presented evidence of at least three types of meaningful between-speaker variability in articulation in addition to the average age- and sex-related differences in vocal apparatus size. Clearly, then, perceptual compensation for speaker differences has to go well beyond mere vocal tract normalization. A promising route for describing how listeners compensate for such variation in speaker style is an instance-based (or exemplar) model of speech perception (Nosofsky, 1986; Kruschke, 1992).

In this kind of model (see figure 6), categories are represented cognitively as exemplars in the psychoacoustic space, a map in which the space covered by any one category is the result of actual perceptual experience. A realistic covering map will have many dimensions, corresponding to the many dimensions of the signal to which the listener attends. However, for convenience, we show only a two-dimensional covering map here, which we exemplify with the first and second formants. That is, each of these squares is a point in the listener's auditory F1-F2 space. Categories, then, are represented by distinct sets of weights which code the strength of association between a location in the psychoacoustic space and a category node. This kind of model can account for robust perception of phonemes as produced by a variety of speakers, and it can also account for robust perception of meaningful speaker categories as speakers produce a variety of phonemes.

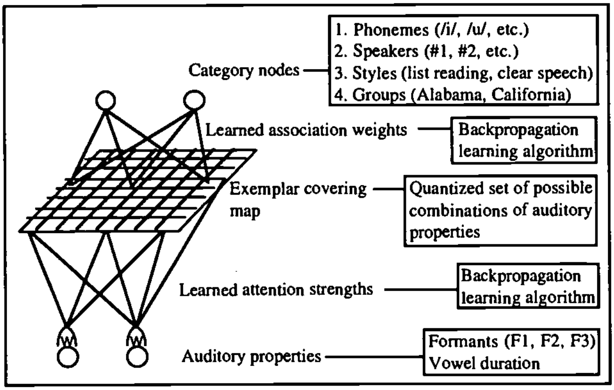

Category nodes
1. Phonemes (/i/, /u/, etc.)
2. Speakers (#1, #2, etc.)
3. Styles (list reading, clear speech)
4. Groups (Alabama, California)
Learned association weights
Backpropagation learning algorithm
Exemplar covering map
Quantized set of possible combinations of auditory properties
Learned attention strengths
Backpropagation learning algorithm
W
W
Auditory properties
Formants (F1, F2, F3)
Vowel duration

Figure 6. A model of speech perception using an exemplar covering map to simultaneously categorize several types of information conveyed by speech.

We used an implementation of the ALCOVE model described by Kruschke (1992). Each token to be categorized was defined by the frequencies of the first three vowel formants at vowel midpoint and by the vowel duration. The covering map was drawn from a set of vowel formant and duration measurements taken from a group of 39 Ohioans (thus the covering map was not representative of either Alabamians or Californians). In calculating the similarity of an input token to the locations in the covering map we used a Euclidian distance measure and a Gausian similarity function. The back-propagation method (Kruschke, 1992) was used to learn both the associations between covering map locations and categories and also the attention strengths given to the stimulus dimensions. Variable parameters in the model, a similarity scale parameter and the attention and association learning rate parameters, were selected by trial and error. With a parameter optimization algorithm we would expect to achieve better vowel classification performance than reported below but no substantial change in the patterns of category structure.

We first trained the model on the utterances produced by the Alabama speakers, a dataset which included between-speaker variation across the 13 female speakers, and also within-speaker variation between normal lab speech and an elicited clear speech style. The model achieved 74% correct vowel classification overall.

Figure 7 shows the association weights between each point in the F1/F2 covering map and the category node for the tense rounded vowel in *who'd*. The open circles plot positive weights, with size scaled to the weight magnitude, and the closed triangles plot negative weights that are substantially less than zero. The weights have a bimodal distribution reflecting the category-internal contrast between the normal lab speech list-reading style and the clear speech style, which had more peripheral F2 values. We've highlighted the two modes here by drawing ellipses around the exemplars with the highest weights. Note also the band of filled triangles just below the ellipses separating the *who'd* category from the *hood* and *owed* categories. These exemplars are the potentially most confusable members of a neighboring category and so are singled out by the training procedure for negative association weights. That is, there is a tuning of the categorization function to sharpen the category boundaries. Note also that there are no such negative weights between the two modes of the *who'd* category weights. In other words, the model does not "normalize" the hyperarticulated clear speech style to convert it to the "normal" style, but represents in the category-internal structure the natural variation actually encountered in the input data. Figure 8 shows the association weights for the vowel in *had* in this model. Again, we have drawn an ellipse to highlight the exemplars with the strongest associations to the category.

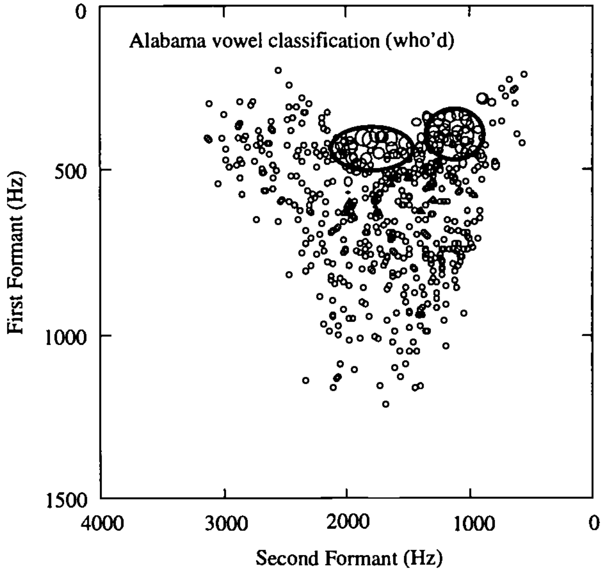


Figure 7. Association weights for who'd in the Alabama data. Each point represents a location in the exemplar covering map. Exemplars which are more strongly associated with *who'd* (large association weights) are given larger points. Points with negative association are plotted with filled triangles.

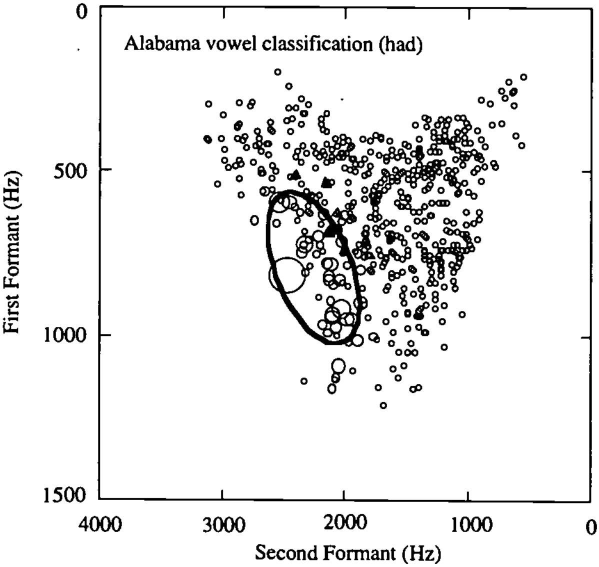


Figure 8. Association for *had* after training on the Alabama data.

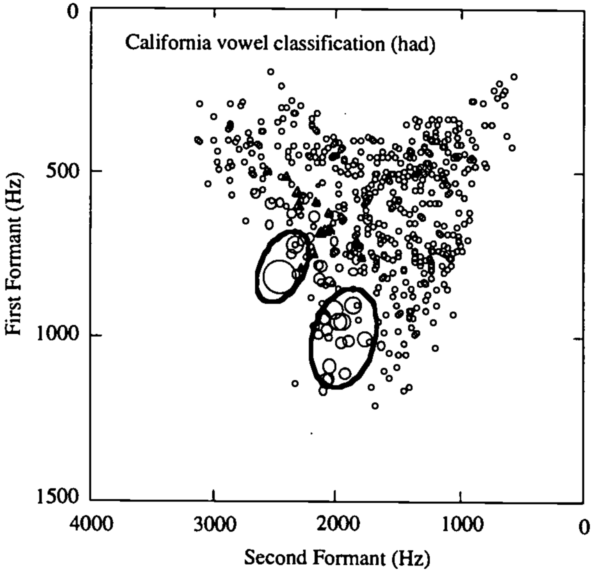


Figure 9. Association weights for *had* after further training on the California data.

We then exposed this Alabama model to utterances produced by the 7 California talkers. At first exposure, classification dropped to only 60% accuracy, but after training, accuracy rose to 77%. Figure 9 shows the association weights for *had* in the elaborated model, which begins to show more category-internal structure, with lower positive values and even some negative weights separating the two modes.

We also trained a model on an orthogonal dimension of classification - to categorize the speaker as either from Alabama or California. The F1/F2 covering map showing the association weights for the category "California speaker" is shown in figure 10. As can be seen in the figure, only tokens with very high F1 values were associated with the category "Californian".

Figure 11 shows the proportion of correctly classified tokens, across the different word types in the corpus. The open bars are for tokens produced by Alabama speakers, and the cross-hatched bars for California speakers.. Since there were more Alabama speakers in the corpus, the model adopted the general strategy of assuming that the speaker was from Alabama in the absence of evidence to the contrary. This, of course, yields 0% correct identification for the California tokens of most words. Words that exemplify vowel categories which have markedly different distributions in the F1-F2 space, however, yield much better dialect classification. In particular, tokens of *had* and *awed* have better than 50% correct classification of the speaker's dialect. The model is sensitive to just those lexical categories which differentiate the two dialects. In other words, just as real listeners do, it hones in on the sociolinguistically meaningful variability in the signal.

We have yet to analyze these corpora for idiosyncratic differences in the relative weighting of F1 and duration for differentiating head from had. Nor do we have data on whether speakers also differentially weight these dimensions in perception. However, studies such as Di Paolo & Faber (1990) and Newman (1996) suggest that there is a relationship. For example, Newman (1996) found a correlation between the average

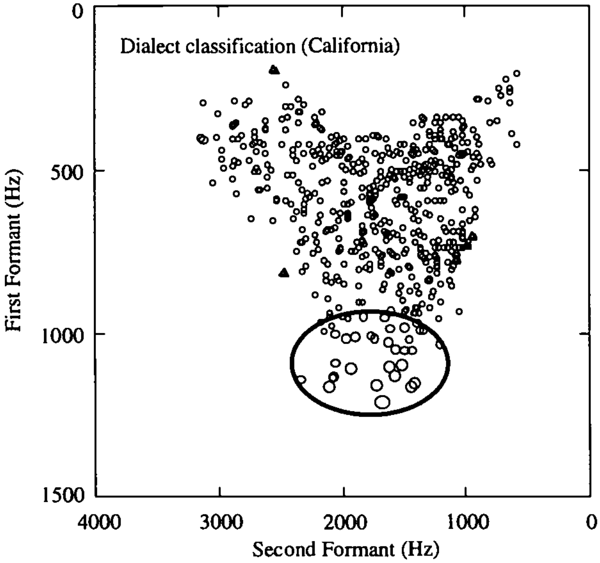


Figure 10. Association weights between locations in the vowel covering map and the speaker category "Californian" in the combined data set.

VOT in subjects' productions of /pa/ and the VOT of synthetic tokens that they rated as the best examples. Di Paolo & Faber (1990) similarly found that younger speakers of Utah English who differentiate the tense vs. lax vowels in *pool* vs. *pull* primarily on the dimension of voice quality rather than F1/F2, also can attend to that "redundant" dimension in categorizing words. We anticipate that an exemplar model can account for such patterns because speaker's own productions and the speech produced in the immediate speech community are likely to be a large component of the exemplar space.

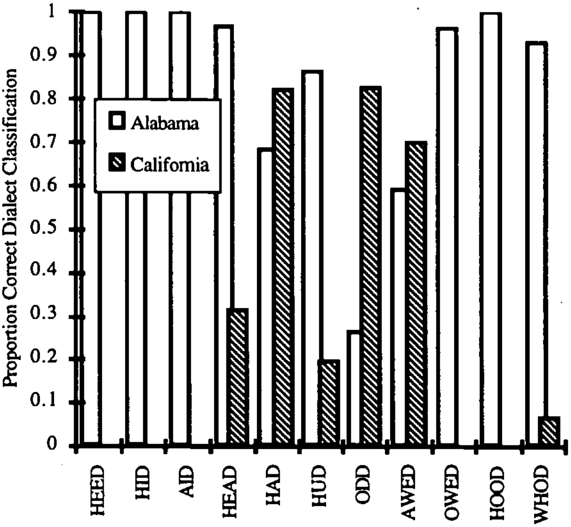


Figure 12. Proportion correct classification of dialect by word.

Conclusions

In this paper we have reviewed evidence for between-speaker differences in speech production that go beyond sex- and age-related differences in physiology. Some of these differences may be related to more subtle morphological differences such as steepness of palate doming. However, there is a substantial residue of arbitrary differences that constitute the speaker's "style". An important component of style differences is the set of differences that can be observed across a well-defined social group, and which indicate group membership. These can be perceptually salient. Thus, perceptual compensation for speaker differences must go beyond mere vocal tract normalization. A promising route for describing how listeners compensate for the arbitrary variation of style is an instance-based (or exemplar) model of speech perception in which the distribution of weights in a covering map are determined by the relative sum of exemplars that the listener encounters for each category. This works for covering maps that let the listener classify the speaker's dialect as well as for covering maps that classify the vowel category.

References

de Jong, K.J. (1995) The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, **97**, 491-504.

Di Paolo, M. & Faber, A. (1990) Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change*, **2**, 155-204.

Edwards, J., Beckman, M.E., & Flechter, J. (1991) Articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, **89**, 369-82.

Harrington, J. & Cassidy, S. (1994) Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English. *Language and Speech*, **37**, 357-73.

Harrington, J. & Fletcher, J. (1996) Acoustic (non)consequences of gestural variability in the production of accentual prominence. *Journal of the Acoustical Society of America*, **100**, 2826-7.

Johnson, K., Flemming, E. & Wright, R. (1993) The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, **69**, 505-28.

Kruschke, J. (1992) ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

Ladefoged, P., DeClerk, J., Lindau, M., & Papcun, G. (1972) An auditory-motor theory of speech production. *UCLA Working Papers in Phonetics*, **22**, 48-75.

Labov, W. (1966) *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Miller, J.D. (1989) Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, **85**, 2114-34.

Nearey, T.M. (1978) *Phonetic Feature Systems for Vowels* (IU Linguistics Club, Bloomington, IN).

Newman, R.S. (1996) Individual differences and the perception-production link *Journal of the Acoustical Society of America*, **99**, 2592.

Nosofsky, R.M. (1986) Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

Peterson, G. & Barney, H. (1952) Control methods used in a study of the identification of vowels. *Journal of the Acoustical Society of America*, **24**, 175-84.

Trudgill, P. (1974) *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.

OSU Working Papers in Linguistics 50 (127-162)

Japanese ToBI Labelling Guidelines*

Jennifer J. Venditti
venditti@ling.ohio-state.edu

1 What is Japanese ToBI?

The Japanese ToBI labelling scheme (J_ToBI) is a method of prosodic transcription for Tokyo Japanese utterances which is consistent with the design principles of the ToBI system for English (see Silverman et al. 1992, Beckman and Hirschberg 1994, and Beckman and Ayers 1994). The purpose of the Japanese ToBI system is to provide a systematic phonological transcription of Japanese prosody which can be used to consistently label corpora at different sites. The J_ToBI system should be able to accurately describe the phonological events in pitch contours of spontaneous speech as well as read lab speech.

A J_ToBI transcription consists of the speech waveform and F0 contour for the utterance and a set of symbolic labels. The mandatory labels are divided into 5 separate label tiers in which labels of the same type are marked: tones, words, break indices, finality and miscellaneous. Other optional user-defined tiers can be added, as appropriate for the focus of research at each particular site. In fact, a separate tier containing the labeller's own comments and flags (e.g. for difficult areas, etc.) is recommended.

The software currently used in making a J_ToBI transcription, and that which is used in the figures in this text, is *Waves+* by Entropic Research Laboratory. However, in theory any speech analysis software may be used, as long as it has the capabilities to align and mark labels on separate tiers. It is also possible to make a J_ToBI transcription in a non-*Waves+* ASCII format (labels and timepoints), as given in Appendix B.

The following sections outline the basics of a J_ToBI transcription and give examples of labelled utterances. This is intended to be a complete and self-contained guide to the prosodic labelling of Japanese, so that anyone with a knowledge of Japanese may be able to use J_ToBI to label their own databases. The entire purpose of Japanese ToBI is to provide a standard for prosodic labelling of diverse speech data. With such a tool in hand, we will be much more prepared to examine current issues in Japanese prosody.

2 Word Tier

The word tier in J_ToBI corresponds to the "orthographic tier" in English ToBI. In this tier, words may be marked using either Japanese orthography or romanization, depending on which one the transcriber is more comfortable with, or which is most appropriate for exporting to relevant applications. The romanized transcription used in the examples in this paper is outlined in Appendix A.

In either type of transcription, lexical accent (where applicable) should be marked on the relevant mora. If labelling words in Japanese orthography, *furigana* should also be provided in order to precisely mark the location of the *akusento kaku*. Accent should be labelled according to the dictionary entry for the word, taking into consideration accent shift or compound formation rules that may apply (usually described in detail at the back of an accent dictionary). Any inconsistencies between the dictionary entry and the speaker's actual production will be marked in the tone tier (see section 3).

What constitutes a minimal "word" in Japanese is a matter of some debate, and will be discussed in more detail below in section 4.2. Filled pauses should also be marked in this tier, and should have some consistent form of transcription, which is agreed upon by labellers at each particular site.

*Copyright 1995 by Jennifer Venditti. This document is an exact reprint of the original "Japanese ToBI Labelling Guidelines" distributed in October 1995.

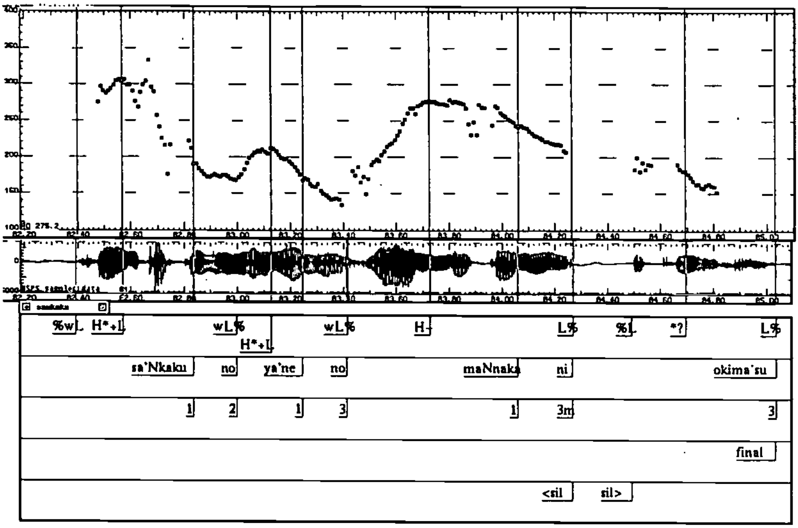


Figure 1: ≪**sankaku**≫ "I will place it right in the center of the triangle roof."

In the word tier, the label for each word is placed at its **right** edge, according to waveform or spectrogram segmentation.

3 Tone Tier

This tier contains the tones of Tokyo Japanese, as in the analysis initially proposed by Beckman and Pierrehumbert (see Beckman and Pierrehumbert 1986, Pierrehumbert and Beckman 1988), and developed further in work by Venditti (see Venditti (forthcoming)). The following subsections describe the inventory of tones for Tokyo Japanese, and give instructions on how to mark them in the J_ToBI tone tier.

3.1 Accent H*+L

This label marks the lexical accent, and should be placed within the accented mora. In many cases, the position of this H*+L accent label will coincide with the location of the actual F0 maximum. However, it is not uncommon to see the peak of the accent (and the fall) occur after the accented mora (e.g. *ososagari*, see Sugito 1981). In such cases, the H*+L label should be placed within the accented mora, and an additional "<" label should be placed at the actual peak to mark the late F0 event (see section 3.3.3 for use of the ">" label in marking an early F0 event). It is essential for the point of F0 maximum to be marked, either with the H*+L label or the < label.

The example utterance ≪sankaku≫ shows the marking of the H*+L pitch accent in cases where the peak occurs within the accented mora. (For now, concentrate on the /sa'Nkaku no ya'ne no/ portion only, and ignore the tiers other than the tone and word tiers. We will return to the other parts of the transcription below. Word-for-word English glosses for each example utterance in this guide are given in Appendix B.) In this utterance, the H*+L accent labels on /sa'Nkaku/ 'triangle' and /ya'ne/ 'roof' are placed at the F0 peak, which falls within the accented mora.

Utterance ≪yane≫ gives an example of an accent peak occurring after the accented mora. The nouns /ya'ne/ 'roof' and /ma'do/ 'window' are initially accented, but clearly the F0 maximum and accentual fall in each word occur well into the second syllable (see the dip in amplitude in the

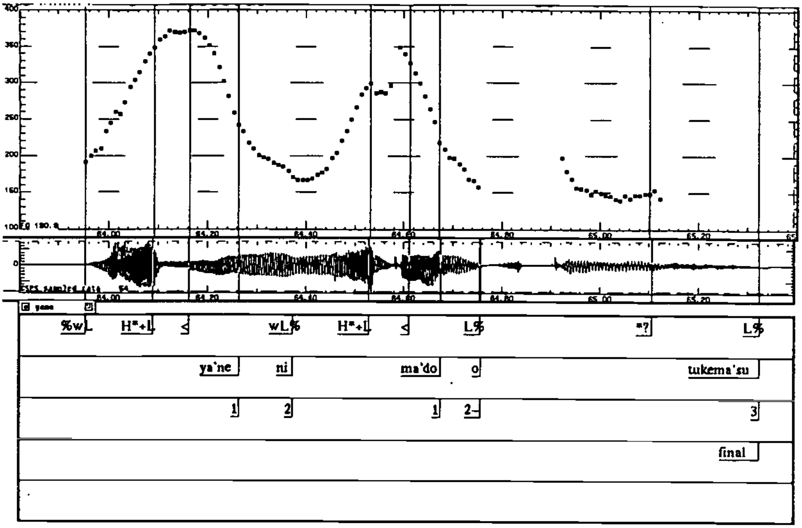


Figure 2: ≪yane≫ "I will put a window on the roof."

waveform for the /n/ of /ya'ne/ and the /d/ of /ma'do/). Here, the H*+L accent label is placed within the accented mora, and the actual F0 peak is marked with the < label (placed at the start of the precipitous fall in pitch).

Labellers should be aware that there may be movements or peaks in the F0 contour which are not genuine tonal events, but rather are segmental perturbations or mistrackings due to creaky voice, etc. For example, the /d/ of /ma'do/ in example utterance ≪yane≫ causes a considerable F0 perturbation. The English ToBI Labelling Guidelines (Beckman and Ayers 1994) provides an extensive discussion of such segmental effects on the F0 contour.

Careful labelling of the actual F0 event in J_ToBI transcribed databases will facilitate future research on the timing of F0 peaks relative to the accented mora (for example, see Sugito 1981, Hata and Hasegawa 1988). Marking the actual high F0 event will also surely help research on the relationship between discourse structure and pitch range or local prominence.

3.1.1 No Mark for Downstep

Unlike in English intonation, where the use of a downstepped accent (!H* or H+!H*, as opposed to H*) is a paradigmatic choice made by the speaker, downstep in Japanese is completely predictable from the lexical accent specification of the preceding phrase. An accentual phrase (whether itself accented or not) will be downstepped if (1) the preceding accentual phrase bears an accent, and (2) both phrases are in the same intonation phrase. (The terms "accentual phrase" and "intonation phrase" are described in more detail below in the sections on tones 3.2, 3.3.1, and sections on break indices 4.3, 4.4.) Downstepping is seen, for example, in the utterance ≪sankaku≫, in which the second phrase /ya'ne no/ 'roof-GEN' is downstepped relative to the preceding accented phrase /sa'Nkaku no/ 'triangle-GEN'. Since the presence or absence of downstep is predictable from the information given in the word tier (i.e. the lexical accentuation of words) and the break index tier (i.e. the type of prosodic boundaries), there is no need to mark downstep in the tone tier of J_ToBI.

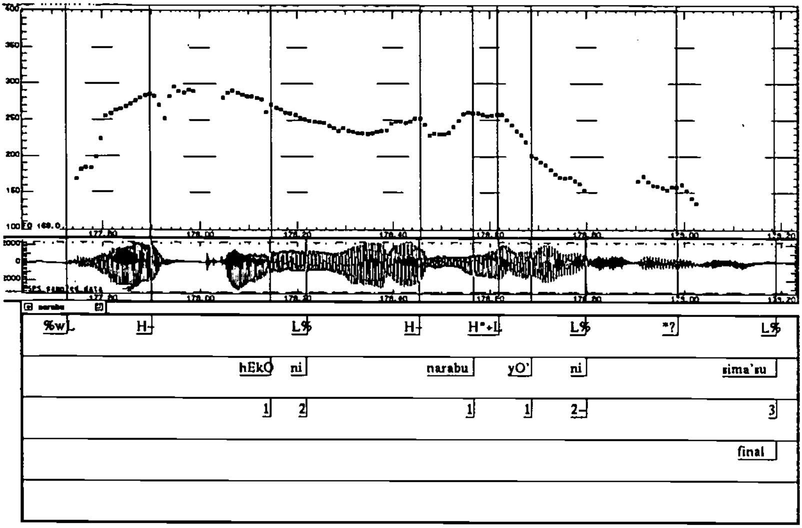


Figure 3: ≪narabu≫"I will make it so that they line up level with each other."

3.2 Phrasal H-

This label marks the H- phrasal tone of the accentual phrase, the level of the prosodic hierarchy above the word. At this level, words may group together into prosodic units delimited by two tones: a H- phrasal tone and L% boundary tone. That is, there is a H- phrasal high near the beginning of the phrase, and a final L% boundary tone marking the end (more on the boundary tone in section 3.3.1). The H- phrasal tone is marked on each unaccented accentual phrase, and on any accented accentual phrase where the H- is distinguishable from of the high tone of the lexical accent (i.e. the shoulder for the H*+L).

The H- label should be placed within the second mora of the phrase. In Tokyo Japanese, this H- phrasal tone is associated phonologically to the second mora. The peak F0 in unaccented phrases (and at the end of the rise in accented phrases where the the H- is distinguishable) should occur around this point. The example ≪narabu≫ shows the marking of this tone in both unaccented and accented phrases. In this utterance, the H- label is marked on the second mora in both phrases. This location coincides with the peak F0 (disregarding segmental perturbations) in the first unaccented phrase /hEkO ni/ 'level-PART', and at the end of the rise in the second accented phrase /narabu yO' ni/ 'line up so that'. Example ≪sankaku≫ also shows the marking of the H- label, placed on the second mora (at the high F0) of the phrase /maNnaka ni/ 'middle-LOC'.

In some utterances the second mora may not coincide with the actual high F0 of the phrase (or the end of the rise in accented phrases). In such cases, the label "<" (late F0 event) should be used to mark the actual high F0 point. This is the same labelling convention used for the H*+L accent label, i.e. the H tone label (H*+L or H-) is placed on the relevant mora, and the < label is used to mark the actual event if it occurs later. The example ≪kazumi≫ shows an utterance in which the high of the phrase is realized after the second mora (after the first word, in fact).

Marking the actual high F0 event associated with the phrasal H- will make it possible to automatically extract an estimate of the pitch range or prominence of unaccented phrases (and even accented phrases in which the H- is higher than the shoulder of the accentual H*+L) for use in future research. Thus, as with H*+L, the labeller should take care to place the H- label (or the < label) at a reliable point in the F0 contour.

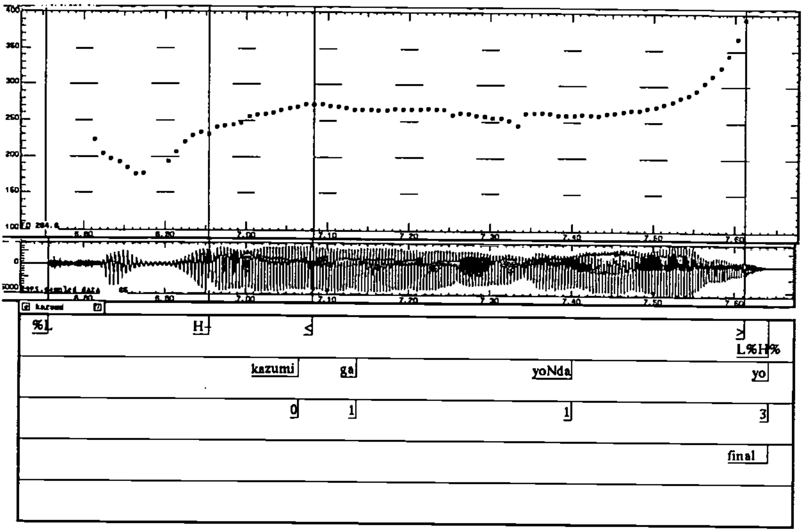


Figure 4: ≪**kazumi**≫ "Kazumi called her."

Words accented on the last mora (e.g. /kami'/ 'paper') which are intonation phrase final do not contain an accentual fall (in Tokyo Japanese), and thus are not distinguishable from unaccented words. In such cases where no fall is observed (e.g. when a pause or some disfluency follows), J_ToBI prescribes that the high F0 of these words be marked using the phrasal H-, and not with the accentual H*+L.

3.3 Boundary Tones

3.3.1 Final L% and wL%

As mentioned in the previous section, the accentual phrase in Tokyo Japanese is delimited by two tones: the H- phrasal tone and a final L% boundary tone. The L% boundary tone is marked in J_ToBI at the right edge of the accentual phrase (see also break index 2 in section 4.3). This L% tone label is aligned with the word and break index labels, and is marked at the end of every accentual phrase, even if there is an additional rise due to a H% or HL% boundary tone (see sections 3.3.3 and 3.3.4 for descriptions of these tones). Example utterance ≪kazumi≫ and others later in this paper show the marking of a L% before the rise to the H%. Example utterances ≪sankaku≫ and ≪yane≫ show the L% boundary tone on utterance final accentual phrases with no rise.

If the immediately following phrase (with no intervening pause) is initially accented, or begins with a long syllable, a wL% ("weak" low) boundary tone is used instead of the L% ("strong" low) tone. In such cases, the L% does not have enough time to be realized fully due to the immediately following high tone, and is undershot, resulting in a weak low (wL%). The utterance ≪sankaku≫ contains two wL% boundary tones marking the edges of the accentual phrases /sa'Nkaku no/ 'triangle-GEN' (occurring before the initially accented word /ya'ne/ 'roof'), and /ya'ne no/ 'roof-GEN' (occurring before a word with a long first syllable /maNnaka/ 'middle'). Example ≪yane≫ also has a wL% boundary tone marking the edge of the first phrase before the initially accented word /ma'do/ 'window'.

3.3.2 Initial %L and %wL

A low boundary tone is also marked at the beginning of utterance initial or post-pausal medial phrases. As with the final low boundary tone, this initial boundary tone has "strong" and "weak" variants depending on the characteristics of the initial part of the phrase, and should be labelled in the same manner as described in section 3.3.1 above. The label should be aligned exactly with the start of phonation, according to the waveform or spectrogram.

Example ≪kazumi≫ shows the marking of %L utterance initially. The first mora of the initial word /kazumi/ is short and bears no accent, hence the use of the "strong" variant of the %L boundary tone. In examples ≪sankaku≫, ≪yane≫ and ≪narabu≫, the utterance initial words have either an accented first mora (≪sankaku≫ and ≪yane≫) or begin with a long syllable (≪narabu≫). In these cases, the "weak" variant of the initial boundary tone, %wL, is used.

This initial low boundary tone provides an anchor from which the F0 rises at the beginning of an utterance or after utterance medial silent intervals. In the case of utterance medial phrases with no preceding pause, the final L% of the preceding phrase serves this purpose (and thus an additional initial %L or %wL boundary tone is not necessary). However, there is one case in which it is necessary to mark an initial low boundary tone even though no silence precedes it. This is when an utterance medial phrase follows a H% boundary tone, with no intervening pause. In such cases an F0 fall is observed from the high boundary tone to a low point at the start of the next phrase. The marking of %wL after H% is shown in the first part of example utterance ≪nibanme≫. (For now, concentrate only on the %wL boundary tone before the word /siNsitu/ 'bedroom'. The rest of this utterance will be discussed in detail below.)

3.3.3 Final H%

This label marks a final high boundary tone for an intonation phrase (see section 4.4 below for discussion of this level of the prosodic hierarchy). This boundary tone typically occurs finally in interrogatives, such as in example ≪nara_quest≫, but is also often found at the end of declarative sentences such as example utterances ≪kazumi≫ and ≪mayumi≫, and also on utterance-medial phrases such as ≪nibanme≫ and others in this guide.

The H% mark should be placed at the right edge of the intonation phrase, aligned exactly with the word and break index marks. In cases where this location does not correspond to the actual maximum value in the F0 contour, labellers should use the early F0 event label (">") to pinpoint the actual event. Before pauses, very often the maximum F0 value of this H% boundary tone will occur before the cessation of phonation (i.e. before the word boundary), due to mistrackings of F0 caused by the rapid decrease in amplitude. The use of the early F0 event label to mark the F0 maximum is shown in example utterances ≪kazumi≫ and ≪mayumi≫.

When marking a H% boundary tone on the right edge of an intonation phrase, labellers should also not forget to label the L% boundary tone of the final accentual phrase (the H- and L% tones delimit each accentual phrase). The combined tone label L%H% (or L%HL%) is thus used for convenience. In example utterances ≪nara_quest≫ and ≪mayumi≫, the rise to the H% from the previous L% is obvious. However, in the utterance ≪kazumi≫, it may not be so apparent. In this utterance, the pitch rises to the H- phrasal tone, then rises a second time to the high boundary tone utterance finally. We know that there is a L% present since the F0 does not continuously rise after the phrasal H-, but is leveled out by the L%, then rises again to the H%.

In addition to utterance-final rises, utterance-medial H% boundary tones also occur, as seen in examples ≪nibanme≫ and ≪pinku_mado≫. In ≪pinku_mado≫, the accentual phrases /pi'Nku no/ 'pink-GEN' and /ma'do o/ 'window-ACC' both form their own intonation phrase (see section 4.4 below), marked by a H% boundary tone and following pause (see Nagahara and Iwasaki 1994 for other examples of utterance-internal high boundary tones).

One last issue that deserves mention regarding the H% boundary tone is the qualitative difference in height between pitch rises. In examples ≪kazumi≫, ≪nara_quest≫, ≪mayumi≫, and ≪pinku_mado≫ we observe fairly high F0 excursions to the top of the phrase's pitch range. However, there are also cases which appear to be qualitatively different from these, in which the pitch rises only part of the way. Compare the utterance in ≪nara_quest≫ to the so-called "insisting" declarative shown in ≪nara_insist≫ (see also the boundary tone on /siNsitu no ma'do wa/ 'bedroom-GEN window-TOP' in example utterance ≪nibanme≫).

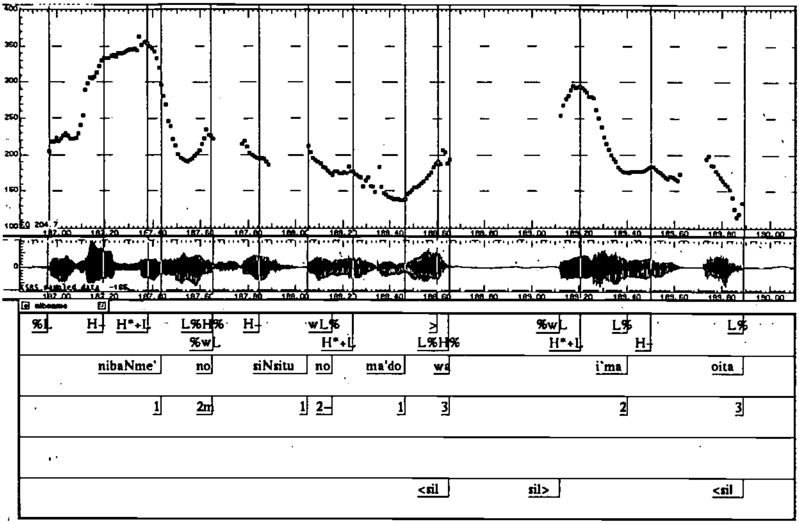


Figure 5: ≪nibanme≫ (part 1) "I will put the second bedroom window below the first window which I just layed down."

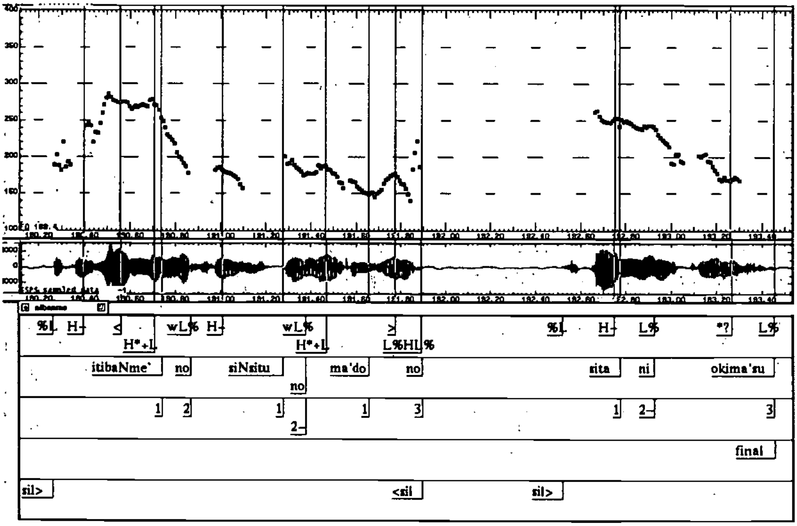


Figure 6: ≪nibanme≫ (part 2) "I will put the second bedroom window below the first window which I just layed down."

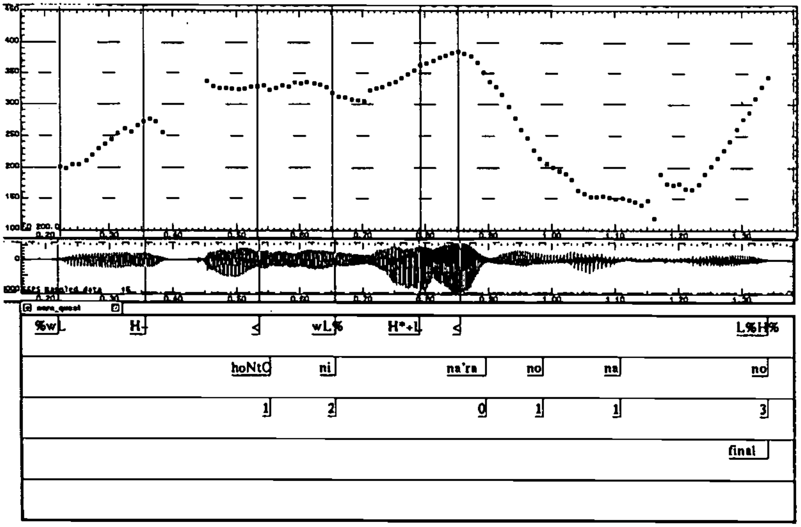


Figure 7: ≪nara_quest≫ "Is that really the one from Nara?"

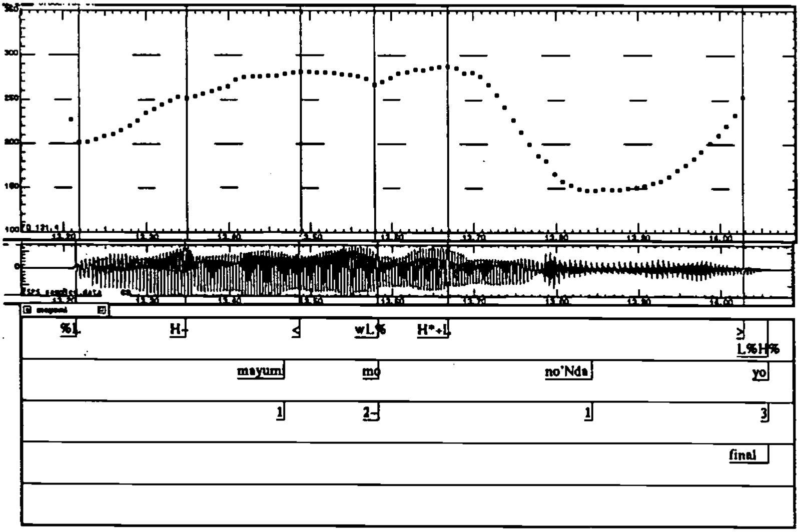


Figure 8: ≪mayumi≫ "Mayumi drank too."

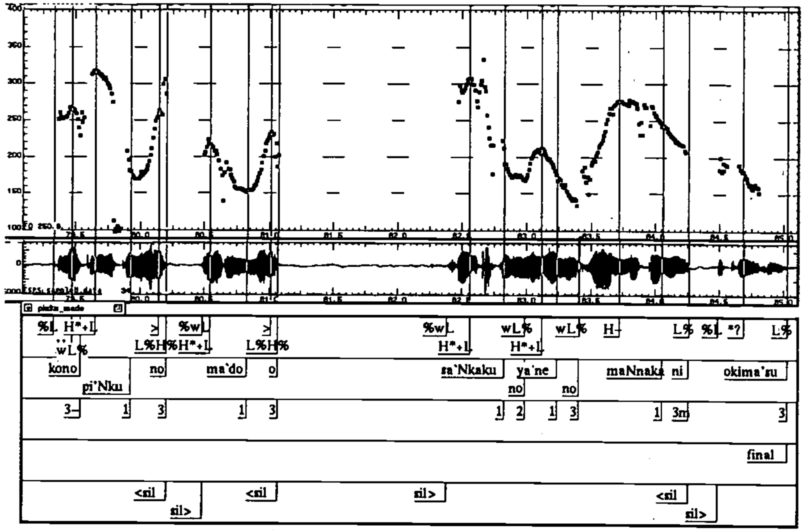


Figure 9: ≪pinku_mado≫ "I will place this pink window right in the center of the triangle roof."

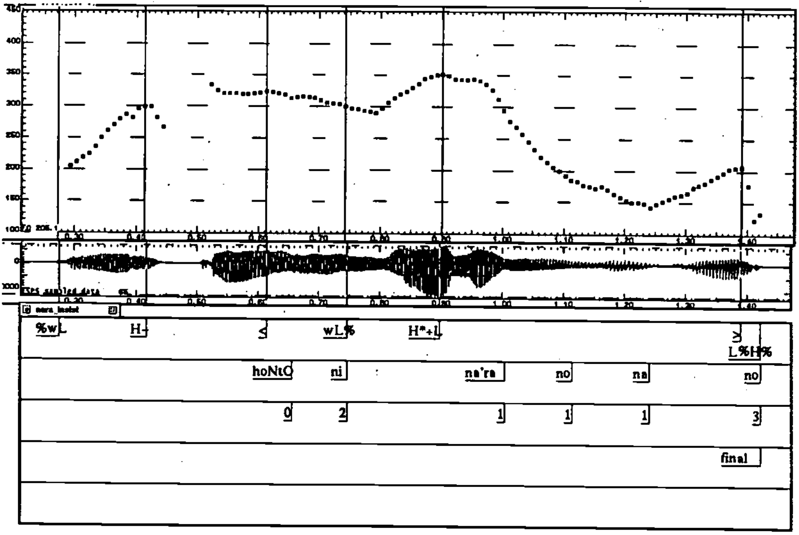


Figure 10: ≪nara_insist≫ "That's really the one from Nara." (insisting)

Since the tone tier in a Japanese ToBI transcription describes only shapes of contours (phonological tonal events), the amplitude of those F0 movements is not documented. There is therefore no way to distinguish between a high-rise and a mid-rise, which are both labelled as H%. For labellers interested in this difference in boundary tone height, a site-specific tier could be added to mark the distinction. This is a good example of how a basic J_ToBI transcription can be expanded to incorporate the research topics of a particular site.

3.3.4 Final HL%

This label marks the high-low boundary tone found that the end of intonation phrases in some speaking styles. Like the H% boundary tone, it should be aligned exactly with the word and break index marks.

The utterance ≪nibanme≫ gives an example of this boundary tone. The F0 contour at the end of the phrase /itibaNme' no siNsitu no ma'do no/ 'first-GEN bedroom-GEN window-GEN' has a marked rise-fall pattern (occurring within the final mora /no/). Here, a L%HL% label is aligned with the edge of this intonation phrase: the L% of the final accentual phrase, and the HL% boundary tone. Note that the ">" early F0 event label is used here to mark the F0 peak of the HL% tone.

3.4 Accent Uncertainty (*?)

As mentioned above, a given string of words may form separate accentual phrases or may group together to form one large accentual phrase. It is most common for unaccented words to group together with adjacent words, while accented words tend to form separate phrases. However, there are cases in which even accented words can join with adjacent words in the same phrase. In such cases, the left-most accented word retains its accentual fall, and the accents to the right in the phrase are totally deleted. Since the difference between a very subordinate accent and deletion of an accent can be subtle, it may be difficult for the labeller to decide whether there is indeed an accent present on a given word (and thus forms a separate phrase), or if the accent has been deleted (i.e. "totally dephrased" with the preceding words). This is especially the case when the pitch range in which the word is realized is very reduced.

In these cases where a word is lexically specified as accented, but upon consideration of both the sound and F0 records the labeller is uncertain whether the speaker indeed produced an accent, a "*?" label should be marked in the tone tier. This label simply means "I don't know if the speaker actually produced an accent", and thus should not be used in cases where the labeller feels that the accent has been totally deleted. Like the H*+L label, *? should be placed within the mora marked for lexical accent.

Utterance ≪narabu≫ shows an example of labeller uncertainty about the accentuation of the final verb /sima'su/ 'do'. It is common in Tokyo Japanese for final predicates to be produced in a reduced pitch range, thus making it difficult to see or hear the accentual fall. It is ambiguous whether the verb has been dephrased together with the preceding words (and thus the accent deleted), or has been produced in a separate phrase with a very narrow pitch range. Here, the *? is marked to indicate this ambiguity. (Section 4.6 will discuss in detail the break index label "2–" which accompanies this tonal uncertainty.)

In example utterance ≪curtain≫, on the other hand, the labeller is not uncertain about the accentuation of the final verb /tukema'su/ 'attach', but rather feels that the accent has been totally deleted. Therefore, no marking (H*+L nor *?) appears on the verb, and the break index (BI 1) indicates that the verb belongs to the same phrase as the preceding words (see section 4.2 below).

The use of the *? uncertainty label is highly subjective, and labeller opinions about whether a word has an accent or not may vary. Since there is no right or wrong answer, labellers should not hesitate to mark *? if they are uncertain. The "degeneration" and total dephrasing of accented words in Japanese is an interesting area of research (see Maekawa 1994), and with these relevant locations flagged by *?, it will then be possible to search through large labelled databases to pull out examples for further research.

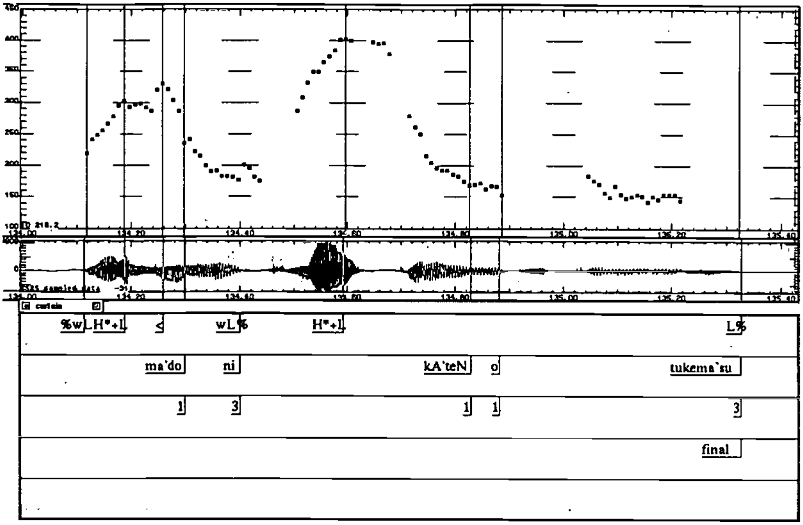


Figure 11: ≪curtain≫ "I will attach a curtain to the window."

4 Break Index Tier

Break indices are labels indicating degree of prosodic association between two sequential units on the word tier. They are markers which show the prosodic grouping of words at various levels. These are subjective values — measures of perceived juncture between adjacent words — and should therefore be labelled upon careful consideration of the sound record. In addition, they will typically have observable physical correlates, such as tonal markings (but see section 4.5 for examples of mismatch). J_ToBI currently distinguishes 4 degrees of disjuncture (on a scale from 0 (weak) to 3 (strong)) in the prosodic structure of Japanese.

All junctures (including filled pauses, cut-off words before restarts, etc.) should be assigned a break index value. The break labels should be aligned exactly with the word labels.

4.1 Break Index 0

This break index marks junctures which are common in fast speech processes, in which there is a very small sense of disjuncture between adjacent words. This may include phenomena such as weakening of velar stops into approximants across word boundaries, or various contracted forms such as /kore+wa/ → [korya] 'this-TOP', /yatte+simau/ → [yattyau] 'do completely', or /no'Nde+iru/ → [no'Nderu] 'is drinking'.

Break index 0 is marked in the example utterance ≪kazumi≫ between the words /kazumi/ 'Kazumi (proper name)' and /ga/ 'NOM'. The velar stop is weakened to an approximant, giving the sense of hardly any separation between the words (there is no audible nasalization of the stop here, which is common for some speakers of the Tokyo dialect).

Example utterance ≪zettai≫ shows break index 0 marked at the boundaries between the verb /kuru/ 'come' and the quotative particle /to/, and between the verb /itte/ 'say' and the perfect progressive marker /ita/. The contracted forms [kurutte] and [itteta] are indicative of the small degree of disjuncture at these boundaries.

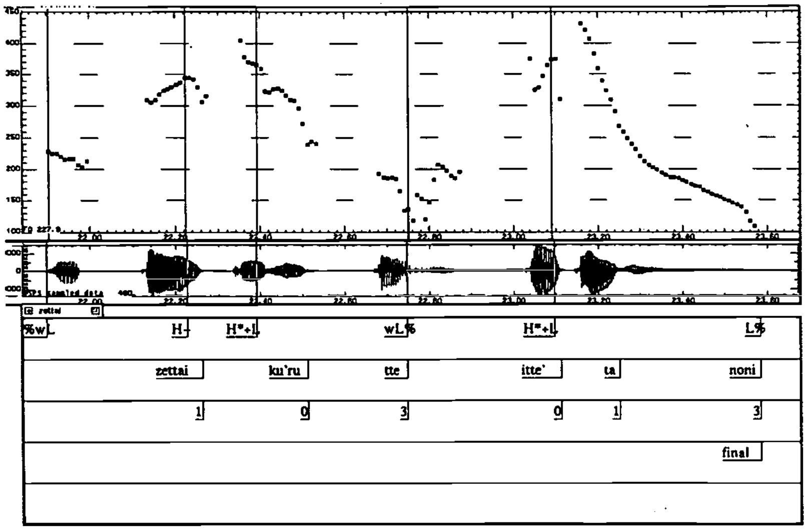


Figure 12: ≪zettai≫ "But she SAID that she would definitely come ..."

4.2 Break Index 1

Break index 1 marks the juncture between two consecutive "words", with no higher-level prosodic boundary. The question of what is a "word" in Japanese is a difficult one, especially concerning the status of postpositions as separate words. J_ToBI does not provide a definitive answer to this question, and currently the dictionary entry (including any inflectional or derivational endings) is taken as the working definition of a "word". Postpositions and sentence particles are treated as separate words in the word and break index tiers. Boundaries marked with a break index 1 have a stronger sense of disjuncture than BI 0, but a smaller disjuncture than BI 2 marking a full accentual phrase break (see section 4.3 below). All of the example utterances shown so far contain examples of the BI 1 label.

4.3 Break Index 2

This break index marks a medium degree of disjuncture between adjacent words. The boundary marked by BI 2 is stronger than that marked by BI 1, but it lacks the cues (e.g. lengthening, pauses, etc.) common to an even stronger boundary marked by BI 3 (see section 4.4).

In most cases the unit marked by a BI 2 at its boundary is characterized by the H- and L% delimiting tones of the accentual phrase, the level of the prosodic hierarchy above the word. However, this perceptually-defined unit (BI 2) and the tonally-defined unit (accentual phrase) are not *always* identical. There may occasionally be cases of mismatch between the perceived juncture and the tonal characteristics (this is described in detail in section 4.5 below).

The utterance ≪sankaku≫ contains a good example of break index 2. The words /sa'Nkaku/ 'triangle' and /no/ 'GEN' are grouped together into one tonally-defined unit (accentual phrase), and the following words /ya'ne/ 'roof' and /no/ 'GEN' into another. The perceived separation between words within each phrase (BI 1) is smaller than the separation between words belonging to adjacent phrases (BI 2). The rise to the H- of the second accentual phrase gives the sense of the beginning of a new unit, and thus we perceive the boundary between the two units as stronger than that within units, or stronger than if there was no phrase break present. In this utterance, the unit marked by BI 2 is identical to the tonally-defined accentual phrase.

Example utterances ≪yane≫ and ≪narabu≫ also show break index 2 marking a medium degree of disjuncture, which in these cases also is identical to an accentual phrase break.

4.4 Break Index 3

Break index 3 marks a strong degree of disjuncture between adjacent words, or between a word and following silent interval. This is the strongest boundary marked in the break index tier of a J_ToBI transcription.

This break index often corresponds to the boundary of the tonally-defined intonation phrase, the highest level of the prosodic hierarchy of Japanese (see Venditti (forthcoming)). The intonation phrase is the prosodic domain within which pitch range is defined and thus within which downstep occurs. At an intonation phrase boundary, the speaker resets to a paradigmatically contrastive new pitch range value for the next phrase. This also is the unit at whose edge a H% (or HL%) boundary tone may occur (see section 3.3.3 above). However, while the unit marked by a break index 3 often corresponds to the intonation phrase, the two are not always identical (see section 4.5 below). As break indices are primarily subjective evaluations of perceived disjuncture, labellers should evaluate the strength of each juncture by carefully considering the sound record, and not only by looking at the F0 contour.

Utterance ≪sankaku≫ also shows an example of the strong disjuncture between adjacent words marked by break index 3. As noted above, this utterance begins with two accentual phrases /sa'Nkaku no/ 'triangle-GEN' and /ya'ne no/ 'roof-GEN'. The boundary between these phrases is marked by BI 2 (medium disjuncture). The boundary between the second phrase /ya'ne no/ and the next phrase beginning with /maNnaka/ 'middle' has an even stronger sense of disjuncture, which is marked with BI 3. Tonally speaking, the first two accentual phrases of the utterance form one larger intonation phrase unit, with downstep causing the second phrase to be lowered with respect to the first (see section 3.1.1 above). The pitch range is then reset on the word /maNnaka/, which is the beginning of the next intonation phrase. The large pitch rise here between intonation phrases gives the sense that a new unit has begun, and thus there is a strong sense of disjuncture between /no/ and /maNnaka/. Example utterances ≪curtain≫ and ≪zettai≫ also show a pitch range reset associated with a strong disjuncture marked by BI 3.

A large F0 rise (e.g. pitch range reset) is one factor that can make labellers sense a strong disjuncture between words. In addition, there may be other factors involved too, including segment lengthening, F0 lowering, decreased amplitude, pauses, etc. The speech signal is full of information that can contribute to the subjective evaluation of disjuncture, and the F0 contour is only one thing.

Example utterances ≪pinku_mado≫ and ≪nibanme≫ (and others later in this guide) show the marking of break index 3 before long pauses. Inserting a pause is one way for a speaker to indicate a separation of information (i.e. a boundary) in the stream of speech. In addition, all of the examples presented here are marked with BI 3 at the end of the utterance. This indicates a strong disjuncture between the final word and the following silent interval (see section 5 below for discussion of marking the degree of finality of these pre-pausal and utterance-final boundaries).

4.5 Mismatch ("m")

All of the example utterances in sections 4.3 and 4.4 above show cases in which the units defined by break indices 2 and 3 corresponded to the prosodic units accentual phrase and intonation phrase, respectively. It was noted that while these perceptually-defined units and tonally-defined units coincide in most cases, there may be instances of mismatch between them, and thus they are not totally redundant to one another. In such cases of mismatch, the labeller should mark the break index according to her/his evaluation of the degree of disjuncture (exactly as in non-mismatch cases), but should then also add the diacritic "m" following the break index value.

Example utterance ≪nibanme≫ shows the marking of 2m. Here, the speaker has produced a H% boundary tone at the edge of the first phrase /nibaNme' no/ 'second-GEN'. As mentioned above in section 4.4, H% and HL% tones are found at the edges of intonation phrases, whose boundaries are most commonly characterized by a strong degree of disjuncture (i.e. BI 3). Yet the disjuncture between the /no/ of the first phrase and the following /siNsitu/ 'bedroom' is clearly not a strong one, but more like a medium disjuncture. The 2m label reflects the fact that the perceived juncture is similar to that commonly associated with an accentual phrase (i.e. BI 2), but that there is a mismatch with the tonal pattern.

The utterance ≪sankaku≫ also shows an example of a mismatch. In this utterance there is a pause between the phrase /maNnaka ni/ 'middle-LOC' and the verb /okima'su/ 'put'. This gives the sense of a strong disjuncture (BI 3). However, the pitch range on the following verb seems reduced, as if the verb has been downstepped. If this is the case, the string /maNnaka ni okima'su/ would form one intonation phrase, within which downstep applies. The mismatch arises from the fact that there appears to be a strong disjuncture marked with BI 3 (which is most often associated with an intonation phrase) *within* an intonation phrase. Therefore, the break has been labelled by 3m to mark this mismatch.

With instances of mismatch flagged using the "m" diacritic, it will enable researchers to search through large labelled databases and investigate these cases further. We will then be in a better position to say how common mismatches are, and whether they are in any way related to certain speaking styles.

4.6 Break Index Uncertainty ("–")

In marking break indices, there may be cases in which the labeller is uncertain about the strength of the juncture. Specifically, it may be difficult to decide between two similar levels, such as 1 and 2. In such cases, the higher break index value should be chosen, and the diacritic "–" should be marked after it. Note that this "–" label does not mean that the strength of the juncture is somewhere in between BI 1 and BI 2, for example. It simply indicates that the labeller is uncertain — (s)he is just not sure of the boundary strength.

Utterance ≪yane≫ shows an example of labeller uncertainty between BI 1 and BI 2. In this utterance, it is not clear if the verb /tukema'su/ 'attach' has been totally dephrased (accent deleted) or if it is just very subordinate and thus realized in a reduced pitch range (see also section 3.4 on accent uncertainty). The break index 2– is used here to indicate this uncertainty about whether the adjacent words are grouped together into a single unit, or remain separate. Since break index 2 often corresponds to the boundary of an accentual phrase (unless there is reason to suspect mismatch), the label 2– indicates uncertainty about the presence of an accentual phrase break. (Note that when the 2– is used, the labeller should also mark the appropriate accentual phrase tones (H- and L% or wL%) in the tone tier.)

Example utterances ≪narabu≫ and ≪nibanme≫ also show the use of 2–. In each of these example utterances, the labeller is uncertain of the boundary strength right before the final verb. Since utterance-final verbs are often realized in a very reduced pitch range, it is ambiguous whether the verb has joined together with the preceding words or forms a separate unit. The break index 2– marks labeller uncertainty due to this ambiguity.

Utterance ≪pinku_mado≫ shows an example of labeller uncertainty between BI 2 and BI 3. Here, the labeller is unsure of the boundary strength between /kono/ 'this' and /pi'Nku/ 'pink'.. The rise in pitch on the word /pi'Nku/ indicates that it may be the start of a new intonation phrase. However, the labeller is uncertain that there is such a strong boundary, and chooses to mark this break with 3–.

As with accent uncertainty described in section 3.4 above, there is no right or wrong answer in using the "–" uncertainty diacritic. It simply allows more freedom to the labeller to express her/his commitment to the break index value assigned. Therefore, labellers should not hesitate use this label liberally. It is only by flagging these uncertain areas that we will be able to go back and take a closer look at them in future research on phrasing.

4.7 Disfluencies ("p")

It is common in spontaneous speech for the speaker to hesitate, stop abruptly and restart, or produce other similar disfluencies. Since the aim of J_ToBI is to describe spontaneous as well as read lab speech, there must be a mechanism for marking such disfluent junctures. Following English ToBI, the diacritic "p" following a break index value is used to mark these cases. The use of this diacritic on the break index tier is a cue that the corresponding tones on the tone tier may be incomplete or ill-formed. Since this "p" label is reserved for disfluent junctures only, labellers should ask

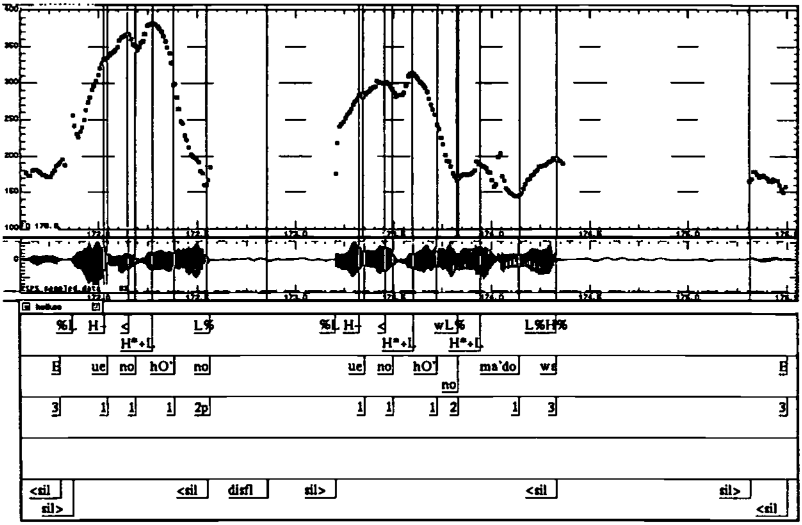


Figure 13: **≪heikoo≫ (Part 1)** "Um, the one on top, the window on top, um, I will make it so that it lines up level with the livingroom window."

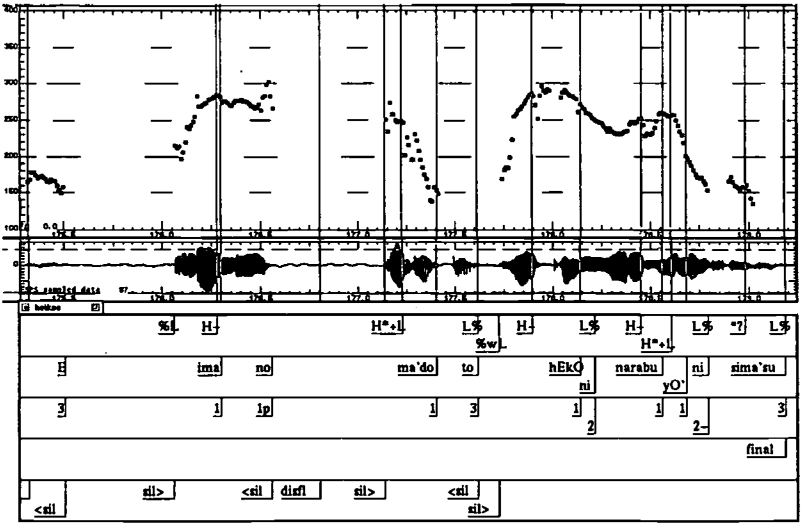


Figure 14: **≪heikoo≫ (Part 2)** "Um, the one on top, the window on top, um, I will make it so that it lines up level with the livingroom window."

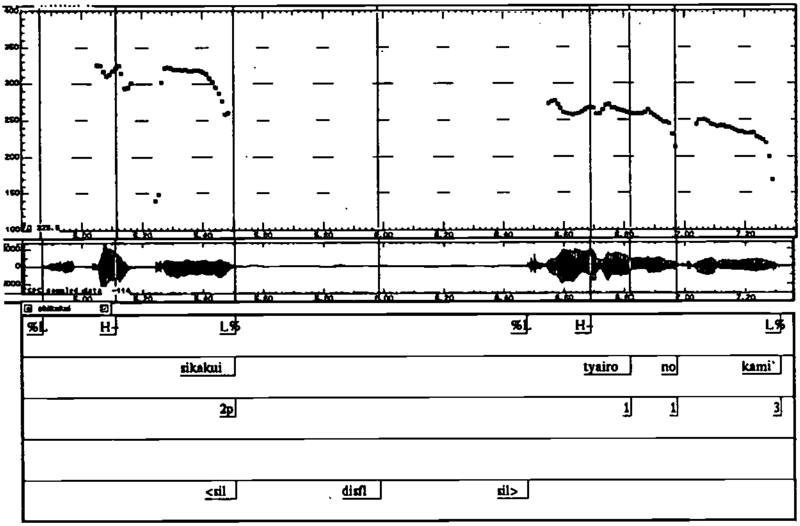


Figure 15: ≪shikakui≫ "The square, brown paper ..."

themselves whether the utterance might have been produced differently (more fluently) if the the speaker was given a second chance to produce it.

A 1p marking on the break index tier indicates cases of abrupt cut-off in which there is no sense of the L% boundary tone which accompanies an accentual phrase juncture (BI 2). Utterance ≪heikoo≫ shows an example of 1p marking. Here, the speaker stops abruptly after the words /ima no/ 'livingroom-GEN' but then continues on with the following /ma'do to/ 'window-with' as if no disfluency had occurred (without restart). Tonally, the string /ima no ma'do to/ constitutes a well-formed accentual phrase (which also happens to be a single intonation phrase). The break index value 1 marked after the /no/ reflects the fact that this juncture falls inside a larger unit (accentual phrase), and the "p" diacritic flags the disfluency (see also section 6 for discussion of the "disfl" label on the miscellaneous tier).

A 2p marking on the break index tier, on the other hand, marks a disfluent juncture which is accompanied by the sense of a L% accentual phrase final boundary tone. Example utterance ≪shikakui≫ shows 2p marking. In this utterance, the speaker hesitates after the word /sikakui/ 'square', but then continues on after a moment with the rest of the phrase /tyairo no kami'/ 'brown-GEN paper'. The downtrend of the words (not downstep here, since the words are unaccented) gives the sense that they are grouped into a single intonation phrase, and that no reset has occurred after the disfluent pause. The boundary after /sikakui/ is a medium disjuncture (BI 2), and indeed if the pause is cut out entirely the utterance sounds like a fluent intonation phrase, with no strong boundary intervening (i.e. BI 3). Therefore, the break index marked here reflects the medium disjuncture (BI 2), as well as the fact that there is a disfluency due to hesitation.

The utterance ≪heikoo≫ also gives an example of break index 2p. There is a disfluent break after the first phrase /ue no hO' no/ 'the one toward the top', and the speaker chooses to restart the utterance after this point. However, the break after /no/ of the first phrase does not have the sense of a strong disjuncture (BI 3), but rather, it sounds as if the speaker would continue on with the utterance, despite the disfluency. Thus, a medium disjuncture is marked (BI 2), along with the "p" diacritic.

Word-internal breaks such as in [tya-tyairo] 'br-brown' should not be indicated on the break index tier, but only by a "disfl" label in the miscellaneous tier (see section 6 below). However, word-internal breaks followed by a restart, such as in [tya- ore'Nzi no kami'] 'the br- orange paper' should be marked using 1p on the break index tier, as well as with a "disfl" label in the miscellaneous tier.

5 Finality Tier

This tier provides a measure of perceived finality of each intonation phrase (break index 3). At present this is a simple binary choice between "final" and "not final". A phrase which is judged as "final" will have at its edge a strong sense of disjuncture, stronger than that of a non-final intonation phrase boundary. This percept of finality should be marked in this tier with the "final" label, which is aligned with the break index label. Those phrases which the labeller perceives as non-final should have no marking.

The notion of "finality" is subjective by nature, and will depend on several acoustic and stylistic factors which, in combination, cue that a phrase is final. The labeller should take into consideration the following phenomena (and possibly others too) when assessing the finality of a phrase:

- final F0 lowering
- segmental lengthening
- creaky voice
- amplitude lowering
- long pauses
- stylized "finality" contours

The labeller should listen the phrase in question and ask her/himself the simple question: "Does the speaker sound done?". If the waveform were to be cut immediately after that break, would it sound as if the speaker had finished her/his turn (or completed an information unit)? Undoubtedly the meaning of the words in the phrase will also play some role in making this judgement, but labellers should concentrate on the sound.

The utterance ≪akete≫ provides an example of finality marking. Here it is the last intonation phrase /sita ni okima'su/ 'below-LOC put' which is marked with the finality label at its right edge. This utterance also provides a good example of the so-called stylized "finality" contour, which is often employed to signal the end of a turn or unit (common in narrative or instructional sequences). In this type of stylized contour, there is typically a H% boundary tone at the edge of the phrase just before the final predicate (note the H% on /akete/ 'open up' here), followed by an optional pause. The final phrase (i.e. predicate) is realized in a very reduced pitch range. This particular combination of tone, pause, and pitch range serves to cue the finality of an utterance.

Another example of finality marking is given in ≪ueshita≫. This utterance shows that fragments, and not just sentence-final phrases containing verbs, can also carry the percept of finality. At three points (aside from the actual end of the utterance) the speaker uses cues such as lowering, lengthening, etc. to signal finality.

Labellers should keep in mind that utterance final intonation phrases are not *always* marked with a final label. In the utterance ≪shikakui≫, the final (and only) intonation phrase is not marked with a "final" label — as the listener will notice, this phrase was cut out of a larger utterance, and thus lacks the finality which is characteristic of end-of-turn phrases.

Sites which choose not to include this a finality tier in the J_ToBI transcription may mark the finality of intonation phrases by a break index 4 on the break index tier. This is essentially equivalent to a BI 3 marking on the break index tier and "final" label on the finality tier. However, we recommend that a separate finality tier be used. We anticipate that marking in this tier will be modified and further developed by sites whose focus is on the various degrees of finality and relationship with discourse structure.

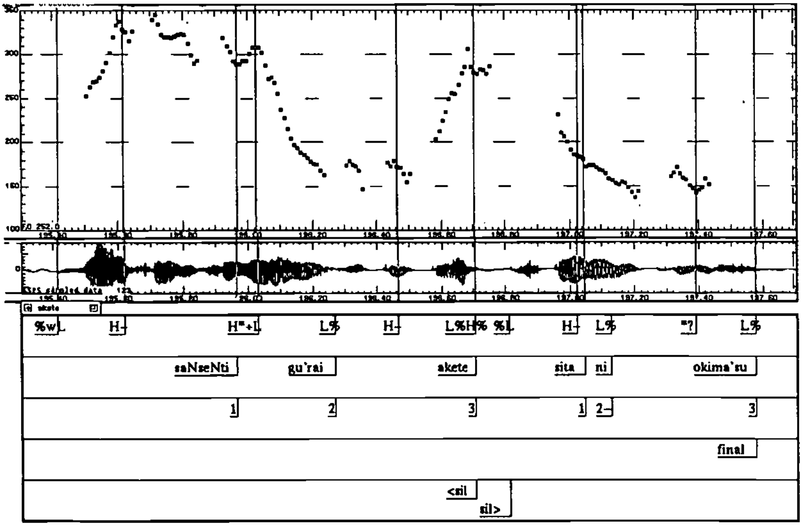


Figure 16: ≪akete≫ "I will open up about a 3cm space and put it below there."

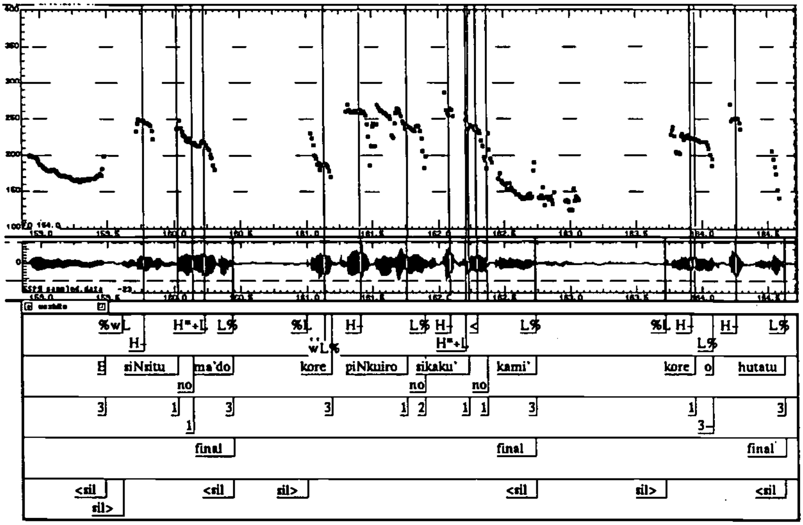


Figure 17: ≪ueshita≫ **(Part 1)** "Uh, the bedroom window, this pink square paper, I will put two of them, line them on top of one another."

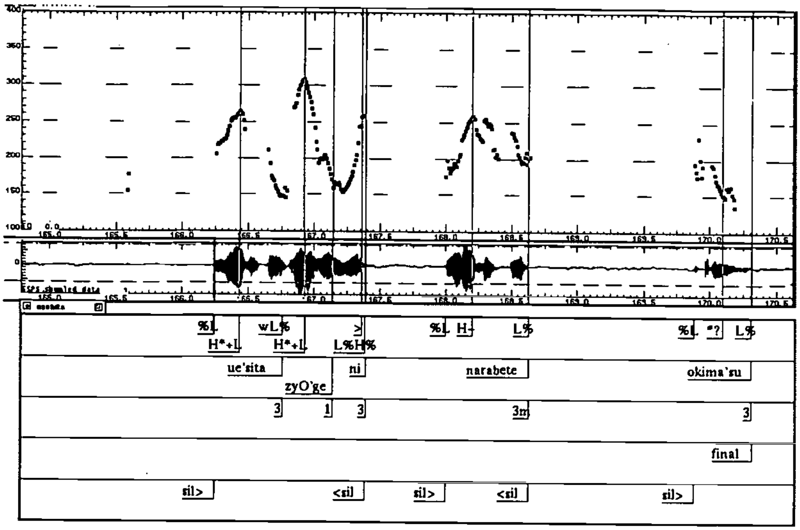


Figure 18: **≪ueshita≫ (Part 2)** “Uh, the bedroom window, this pink square paper, I will put two of them, line them on top of one another.”

6 Miscellaneous Tier

This tier is reserved for other phenomena present in the speech signal which cannot be properly described by the phonological events marked in the tone and break index tiers. Such phenomena include repairs, disfluencies, silences, laughing, etc. Those phenomena which span clearly defined intervals (such as silences, laughing, etc.) should be marked by a pair of labels at their temporal beginnings and ends (e.g. `laugh< .... laugh>`). Other phenomena which are less clearly defined temporally should be marked by a single flag (e.g. “disfl”) at the approximate location. Since disfluencies often effect the corresponding tonal events and break index marks, labels in the miscellaneous tier will serve as flags to identify possible unfinished or ill-formed sequences on the other tiers.

Labeller comments may also be included in the miscellaneous tier, or additional tiers can be added to fit the needs of research at each particular site.

7 Online Data Files and Future Versions

As more speech data becomes available, the Japanese ToBI labelling guidelines may be reformulated and refined. Please check the following World Wide Web pages for revised versions:

`http://ling.ohio-state.edu/Phonetics/J_ToBI/jtobi_homepage.html`
`http://www.itl.atr.co.jp/ToBI/jtobi.html`

A postscript version of this guide and example utterances are also available on these WWW pages. This allows non-*Waves+* users to have access to the sound and F0 files. In addition, there is also a link to the ftp site containing the *Waves+* formatted data files.

If you would like to be placed on the J_ToBI mailing list to receive notices of updates and participate in discussion of J_ToBI labelled utterances, please follow these steps:

1) send e-mail to: `majordomo@ling.ohio-state.edu`
2) body of the message should say only: `subscribe jtobi`
(the contents of the subject line do not matter)

Other comments or questions concerning J_ToBI are also welcome at:
venditti@ling.ohio-state.edu

Acknowledgements

The work reported in this paper was supported in part by a Title VI Foreign Language Area Studies Fellowship, and by a university relations summer fellowship from AT&T Bell Telephone Laboratories. This guide was completed while I was a visitor at ATR Interpreting Telecommunications Research Laboratories and Osaka Gakuin University. I would like to thank Mary Beckman, Nick Campbell, Toshio Hirai, Sun-Ah Jun, Y. Ohta, Marc Swerts, Junko Tsumaki, and Shigemi Yamato for their helpful discussions and insights. However, all mistakes are my own.

References

Beckman, M. E., and G. M. Ayers. 1994. Guidelines for ToBI Labelling. Unpublished manuscript, Ohio State University. [Send e-mail to tobi@ling.ohio-state.edu for ordering information, or visit the English ToBI homepage at http://ling.ohio-state.edu/Phonetics/etobi_homepage.html].

Beckman, M. E., and J. Hirschberg. 1994. The ToBI Annotation Conventions. Unpublished manuscript, Ohio State University and AT&T Bell Telephone Laboratories.

Beckman, M. E., and J. B. Pierrehumbert. 1986. Intonational Structure in Japanese and English. *Phonology Yearbook* 3:255–309.

Hata, K., and Y. Hasegawa. 1988. Delayed pitch fall phenomenon in Japanese. In *Proceedings of the Western Conference on Formal Linguistics*, 87–100.

Maekawa, K. 1994. Is there 'dephrasing' of the accentual phrase in Japanese? In *Working Papers in Linguistics: Papers from the Linguistics Laboratory*, ed. J. J. Venditti, Vol. 44, 146–165. Ohio State University.

Nagahara, H., and S. Iwasaki. 1994. Tail pitch movement and the intermediate phrase in Japanese. Paper presented at the Linguistic Society of America annual meeting, January 1994.

Pierrehumbert, J. B., and M. E. Beckman. 1988. *Japanese Tone Structure*. MIT Press.

Silverman, K. E. A., M. Beckman, J. F. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. TOBI: A standard for Labeling English Prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, Vol. 2, 867–870. Banff, Canada.

Sugito, M. 1981. Timing relationship between articulation and F0 lowering for word accent. *Gengo kenkyuu* 77.

Venditti, J. J. (forthcoming). Developments in a theory of Japanese intonation: From *Japanese Tone Structure* to J_ToBI. Unpublished manuscript, Ohio State University.

Appendix A Romanization

The following specifies how to transcribe words on the word tier with the romanization system used in the examples in this paper. It is only slightly different from the phonemic *kunreisiki* style of romanization. Accent location is marked by an apostrophe (') after the accented mora.

CONSONANTS

p	/patiNko/	'pachinko'
b	/basyo/	'place'
t	/tabe'ru/	'eat'
	/ti'zu/	'map'
	/tugi'/	'next'
d	/dame'/	'useless'
k	/kagi'/	'key'
g	/ga'maN/	'perseverance'
h	/ha'ru/	'spring'
	/hito'tu/	'one'
	/huzi/	'wisteria'
s	/sake/	'sake'
	/siro'i/	'white'
z	/zabu'toN/	'sitting cushion'
	/zibuN/	'one's self'
	/tezu'kuri/	'hand-made'
r	/raineN/	'next year'
m	/makeru/	'lose'
n	/na'be/	'pot'
N	/oNseN/	'hot spring'
y	/yawaraka'i/	'soft'
w	/waka'ru/	'understand'

pp	/kappukE'ki/	'cupcake'
tt	/waratta/	'laughed'
kk	/mikka/	'three days'
dd	/be'ddo/	'bed'
ss	/massu'gu/	'straight'
	/massi'ro/	'pure white'
Nn	/koNnitiwa'/	'Hello.'

py	/happyaku'/	'eight hundred'
by	/byOiN/	'hospital'
ty	/tyairo/	'brown'
ky	/okyakusan/	'guest'
gy	/gyaku/	'opposite'
hy	/hyaku'/	'one hundred'
sy	/sya'mozi/	'rice paddle'
zy	/zyama/	'hinderance'
ry	/ryukkusa'kku/	'backpack'
my	/myO'zi/	'surname'
ny	/nyU'su/	'news'

VOWELS

a	/ka'ge/	'shadow'
i	/inu'/	'dog'
u	/ue/	'top'
e	/eho'N/	'picture book'
o	/koke'/	'moss'
A	/bAgeNsE'ru/	'bargain sale'
I	/kawaI'/	'cute'
U	/kU'ki/	'air'
E	/seNsE'/	'teacher'
O	/hOritu/	'law'

Identical vowels belonging to adjacent syllables are distinguished from long vowels by being written with two sequential letters, as in:

/baai/	'occasion'
/midoriiro/	'green'
/kooni/	'imp'

Diphthongs are distinguished from otherwise identical sequences of heterosyllabic vowels by putting a hyphen between the vowels in the heterosyllabic sequence, as in:

/ko-inu/	'puppy'	(heterosyllabic)
/ko'i/	'love'	(diphthong)
/kusa-iro/	'(darkish) green'	(heterosyllabic)
/kusa'i/	'stinking'	(diphthong)
/usu-iro/	'pale (of color)'	(heterosyllabic)
/su'iro/	'water conduit'	(diphthong)

Appendix B Non-*Waves+* J_ToBI transcriptions

The following are the Non-*Waves+* ASCII versions of the J_ToBI transcriptions for utterances shown in this paper in order of appearance. Non-*Waves+* transcriptions are divided into the following fields:

1 words
2 ^tones
3 $break index
4 @BI timepoint
5 &finality
6 ;miscellaneous
7 #English gloss

```
-----------------------------------------------------------------------------------
<<sankaku>>
sa'Nkaku    ^%wL H*+L    $1    @82.836998    &       ;                #triangle
no          ^wL%         $2    @82.999528    &       ;                #GEN
ya'ns       ^H*+L        $1    @83.247598    &       ;                #roof
no          ^wL%         $3    @83.415830    &       ;                #GEN
maNnaka     ^H-          $1    @84.063096    &       ;                #middle
ni          ^L%          $3m   @84.271248    &       ;<sil            #LOC
                                                     ;sil> 84.496926
okima'su    ^%L *? L%    $3    @85.041123    &final  ;                #put
-----------------------------------------------------------------------------------
<<yane>>
ya'ne       ^%wL H*+L <  $1    @64.264530    &       ;                #roof
ni          ^wL%         $2    @64.373958    &       ;                #LOC
ma'do       ^H*+L <      $1    @64.676112    &       ;                #window
o           ^L%          $2-   @64.757775    &       ;                #ACC
tukema'su   ^*? L%       $3    @65.326151    &final  ;                #attach
-----------------------------------------------------------------------------------
<<narabu>>
hEkO        ^%wL H-      $1    @78.143512    &       ;                #level
ni          ^L%          $2    @78.218332    &       ;                #PART
narabu      ^H-          $1    @78.566405    &       ;                #line up
yO'         ^H*+L        $1    @78.686767    &       ;                #so that
ni          ^L%          $2-   @78.800623    &       ;                #PART
sima'su     ^*? L%       $3    @79.190986    &final  ;                #do
-----------------------------------------------------------------------------------
<<kazumi>>
kazumi      ^%L H-       $0    @7.066480     &       ;                #(name)
ga          ^<           $1    @7.138377     &       ;                #NOM
yoNda       ^            $1    @7.402000     &       ;                #called
yo          ^> L%H%      $3    @7.643574     &final  ;                #SEN_PART
-----------------------------------------------------------------------------------
<<nibanms>>
nibaNms'    ^%L H- H*+L  $1    @187.437953   &       ;                #second
no          ^L%H%        $2m   @187.653405   &       ;                #GEN
siNsitu     ^%wL H-      $1    @188.053084   &       ;                #bedroom
no          ^wL%         $2-   @188.159249   &       ;                #GEN
ma'do       ^H*+L        $1    @188.468375   &       ;                #window
wa          ^> L%H%      $3    @188.655725   &       ;<sil            #TOP
                                                     ;sil>189.122388
i'ma        ^%wL H*+L L% $2    @189.410530   &       ;                #now
oita        ^H- L%       $3    @189.896873   &       ;<sil            #put
                                                     ;sil>190.258653
itibaNme'   ^%L H-< H*+L $1    @190.741026   &       ;                #first
no          ^wL%         $2    @190.869560   &       ;                #GEN
```

```
siNsitu    ^H-          $1   @191.276004  &      ;             #bedroom
no         ^wL%         $2-  @191.376747  &      ;             #GEN
ma'do      ^H*+L        $1   @191.658131  &      ;             #window
no         ^> L%HL%     $3   @191.897829  &      ;<sil         #GEN
                                                 ;sil>192.524955
sita       ^%L H-       $1   @192.781298  &      ;             #below
ni         ^L%          $2-  @192.929732  &      ;             #LOC
okima'su   ^*? L%       $3   @193.458247  &final ;             #put
-----------------------------------------------------------------------
<<nara_quest>>
hoNtO      ^%wL H- <    $1   @0.553271    &      ;             #really
ni         ^wL%         $2   @0.657389    &      ;             #PART
na'ra      ^H*+L <      $0   @0.896246    &      ;             #(pl. name)
no         ^            $1   @0.986889    &      ;             #GEN
na         ^            $1   @1.109380    &      ;             #COP
no         ^L%H%        $3   @1.342113    &final ;             #SEN_PART
-----------------------------------------------------------------------
<<mayumi>>
mayumi     ^%L H-       $1   @13.466608   &      ;             #(name)
mo         ^< wL%       $2-  @13.582692   &      ;             #also
no'Nda     ^H*+L        $1   @13.844117   &      ;             #drank
yo         ^> L%H%      $3   @14.058353   &final ;             #SEN_PART
-----------------------------------------------------------------------
<<pinku_mado>>
kono       ^%L H- wL%   $3-  @79.540736   &      ;             #this
pi'Nku     ^H*+L        $1   @79.927887   &      ;             #pink
no         ^> L%H%      $3   @80.206763   &      ;<sil         #GEN
                                                 ;sil>80.490035
ma'do      ^%wL H*+L    $1   @80.834235   &      ;             #window
o          ^> L%H%      $3   @81.072629   &      ;<sil         #ACC
                                                 ;sil>82.392984
sa'Nkaku   ^%wL H*+L    $1   @82.836998   &      ;             #triangle
no         ^wL%         $2   @82.999528   &      ;             #GEN
ya'ne      ^H*+L        $1   @83.247598   &      ;             #roof
no         ^wL%         $3   @83.415830   &      ;             #GEN
maNnaka    ^H-          $1   @84.063096   &      ;             #middle
ni         ^L%          $3m  @84.271248   &      ;<sil         #LOC
                                                 ;sil>84.496926
okima'su   ^%L *? L%    $3   @85.041123   &final ;             #put
-----------------------------------------------------------------------
<<nara_insist>>
hoNtO      ^%wL H- <    $0   @0.656519    &      ;             #really
ni         ^wL%         $2   @0.747603    &      ;             #PART
na'ra      ^H*+L        $1   @1.005674    &      ;             #(pl. name)
no         ^            $1   @1.115734    &      ;             #GEN
na         ^            $1   @1.242239    &      ;             #COP
no         ^> L%H%      $3   @1.424407    &final ;             #SEN_PART
-----------------------------------------------------------------------
<<curtain>>
ma'do      ^%wL H*+L <  $1   @134.299321  &      ;             #window
ni         ^wL%         $3   @134.401580  &      ;             #LOC
kA'teN     ^H*+L        $1   @134.830093  &      ;             #curtain
o          ^            $1   @134.883657  &      ;             #ACC
tukema'su  ^L%          $3   @135.330491  &final ;             #attach
-----------------------------------------------------------------------
<<zettai>>
zettai     ^%wL H-      $1   @22.262873   &      ;             #absolutely
```

```
ku'ru     ^H*+L        $0   @22.512168    &      ;              #come
tte       ^wL%         $3   @22.750384    &      ;              #QUOT
itte'     ^H*+L        $0   @23.114171    &      ;              #said
ta        ^            $1   @23.252668    &      ;              #PROGRESS
noni      ^L%          $3   @23.586909    &final ;              #PART
-------------------------------------------------------------------------------
<<heikoo>>
E         ^            $3   @171.803163   &      ;<sil          #um
                                                 ;sil>171.865027
ue        ^%L H-       $1   @172.043665   &      ;              #upper
no        ^<           $1   @172.190412   &      ;              #GEN
hO'       ^H*+L        $1   @172.381998   &      ;              #
no        ^L%          $2p  @172.565247   &      ;<sil          #GEN
                                                 ;disfl
                                                 ;sil>173.203382
ue        ^%L H-       $1   @173.352157   &      ;              #upper
no        ^<           $1   @173.498904   &      ;              #GEN
hO'       ^H*+L        $1   @173.727177   &      ;              #
no        ^wL%         $2   @173.833161   &      ;              #GEN
ma'do     ^H*+L        $1   @174.142959   &      ;              #window
wa        ^L%H%        $3   @174.334545   &      ;<sil          #TOP
                                                 ;sil>175.322155
E         ^            $3   @175.506354   &      ;<sil          #um
                                                 ;sil>176.060629
ima       ^%L H-       $1   @176.299048   &      ;              #livingroom
no        ^            $1p  @176.559290   &      ;<sil          #GEN
                                                 ;disfl
                                                 ;sil>177.139433
ma'do     ^H*+L        $1   @177.405076   &      ;              #window
to        ^L%          $3   @177.619776   &      ;<sil          #with
                                                 ;sil>177.727658
hEkO      ^%wL H-      $1   @178.143512   &      ;              #level
ni        ^L%          $2   @178.218332   &      ;              #ADV
narabu    ^H-          $1   @178.566405   &      ;              #line up
yO'       ^H*+L        $1   @178.686767   &      ;              #so that
ni        ^L%          $2-  @178.800623   &      ;              #PART
sima'su   ^*? L%       $3   @179.190986   &final ;              #do
-------------------------------------------------------------------------------
<<shikakui>>
sikakui   ^%L H- L%    $2p  @5.502376     &      ;<sil          #square
                                                 ;disfl
                                                 ;sil>6.477005
tyairo    ^%L H-       $1   @6.820425     &      ;              #brown
no        ^            $1   @6.976449     &      ;              #GEN
kami'     ^L%          $3   @7.319700     &      ;              #paper
-------------------------------------------------------------------------------
<<akete>>
saNseNti  ^%wL H-      $1   @195.964593   &      ;              #3cm
gu'rai    ^H*+L L%     $2   @196.270458   &      ;              #about
akete     ^H- L%H%     $3   @196.707749   &      ;<sil          #open up
                                                 ;sil>196.814808
sita      ^%L H-       $1   @197.049458   &      ;              #below
ni        ^L%          $2-  @197.130703   &      ;              #LOC
okima'su  ^*? L%       $3   @197.579943   &final ;              #put
-------------------------------------------------------------------------------
<<ueshita>>
E         ^            $3   @159.476999   &      ;<sil          #um
```

```
                                                 ;sil>159.605087
siNsitu    ^%wL H-     $1    @160.025312  &      ;               #bedroom
no         ^           $1    @160.141215  &      ;               #GEN
ma'do      ^H*+L L%    $3    @160.435432  &final ;<sil           #window
                                                 ;sil>160.997869
kore       ^%L H- wL%  $3    @161.188805  &      ;               #this
piNkuiro   ^H-         $1    @161.763865  &      ;               #pink
no         ^L%         $2    @161.897600  &      ;               #GEN
sikaku'    ^H- H*+L    $1    @162.236395  &      ;               #square
no         ^<          $1    @162.379046  &      ;               #GEN
kami'      ^L%         $3    @162.740130  &final ;<sil           #paper
                                                 ;sil>163.726736
kore       ^%L H-      $1    @163.956116  &      ;               #this
o          ^L%         $3-   @164.088244  &      ;               #ACC
hutatu     ^H- L%      $3    @164.633626  &final ;<sil           #two
                                                 ;sil>166.243984
ue'sita    ^%L H*+L wL% $3   @166.764472  &      ;               #above/below
zyO'ge     ^H*+L       $1    @167.145487  &      ;               #above/below
ni         ^> L%H%     $3    @167.393008  &      ;<sil           #PART
                                                 ;sil>168.002199 #
narabete   ^%L H- L%   $3m   @168.636177  &      ;<sil           #line up
                                                 ;sil>169.890971
okima'su   ^%L *? L%   $3    @170.320860  &final ;               #put
```

Appendix C J_ToBI Labelling Conventions

This Appendix is intended to serve as a reference to the J_ToBI labels and conventions introduced in this guide, and therefore should be consulted only after the labeller has reviewed the examples and explanations in the preceding sections.

Synopsis

The Japanese ToBI labelling scheme (J_ToBI) is a method of prosodic transcription for Tokyo Japanese utterances which is consistent with the design principles of the ToBI system for English (see Silverman et al. 1992, Beckman and Hirschberg 1994, and Beckman and Ayers 1994). The purpose of the Japanese ToBI system is to provide a systematic phonological transcription of Japanese prosody which can be used to consistently label corpora at different sites.

A J_ToBI transcription consists of the speech waveform and F0 contour for the utterance and a set of symbolic labels. The mandatory labels are divided into 5 separate label tiers in which labels of the same type are marked: tones, words, break indices, finality and miscellaneous. Other optional user-defined tiers can be added, as appropriate for the focus of research at each particular site. In fact, a separate tier containing the labeller's own comments and flags (e.g. for difficult areas, etc.) is recommended.

Word Tier

This tier contains the individual words of the utterance, transcribed using either Japanese orthography or some conventional romanization. Word labels should be marked at the right edge of each word, according to waveform or spectrogram segmentation. We currently take a minimal dictionary entry as the working definition of a "word", and as such we mark postpositions and particles as separate words. Accented words are marked with an apostrophe (') after the relevant mora.

Tone Tier

This tier contains the tones of Tokyo Japanese, as in the analysis initially proposed by Beckman and Pierrehumbert (see Beckman and Pierrehumbert 1986, Pierrehumbert and Beckman 1988), and developed further in work by Venditti (see Venditti (forthcoming)). The J_ToBI tone labels can be divided into three groups: those concerning lexical accent, accentual phrase tones, and intonation phrase boundary tones.

Lexcial Accent

H*+L This bitonal pitch accent marks the lexically specified accent of accented phrases. This label should be placed within the accented mora. In cases in which the F0 peak occurs after the accented mora (e.g. *ososagari*), an additional label "<" (late F0 event) should be placed at the actual F0 peak.

***?** This accent uncertainty label marks labeller uncertainty about whether an accentual fall is present on a word marked as accented in the lexicon. This label is commonly used in phrases with a narrow pitch range (e.g. utterance-final verbs), where it is difficult to distinguish a very subordinate accent from a totally deleted accent.

Accentual Phrase

H- This phrasal high tone is marked on the second mora in unaccented phrases, and also in accented phrases in which the H- is distinguishable from the shoulder of the H*+L accent. It is one of the two tones which delimit the accentual phrase in Japanese (see also break index 2). In cases in which the F0 peak in unaccented phrases (or the end of the rise in accented phrases) occurs after the second mora, the label "<" (late F0 event) should be used to mark the actual high F0 point.

L% Along with the phrasal H-, this final low boundary tone characterizes the accentual phrase. It should be placed at the right edge of each phrase, aligned with the word and break index labels. There is also a "weak" variant of this tone (wL%) used in cases where the next phrase (with no intervening pause) begins with a long syllable, or is initially accented.

%L This intial low boundary tone is marked at the beginning of post-pausal phrases. It provides an anchor from which the F0 rises at the beginning of utterances and after pauses. As with the final low boundary tone, this initial tone also has a "weak" variant (%wL), used in the same contexts.

Intonation Phrase

H% This intonation phrase final high boundary tone marks the final rise common in interrogative utterances and also in some declaratives. It should be marked at the right edge of the intonation phrase (see also break index 3), aligned exactly with the word and break index labels. In cases in which the high F0 at the end of the rise occurs before the right edge of the phrase, an additional label ">" (early F0 event) should be used to mark the actual high F0 point.

HL% This final high-low boundary tone marks the rise-fall contour found at the end of some intonation phrases. This label should also be placed exactly at the right edge of the phrase, and the actual F0 peak should be marked using the ">" early F0 event label.

Break Index Tier

Break indices are labels indicating degree of prosodic association between two sequential units on the word tier. They are markers which show the prosodic grouping of words at various levels. These are subjective values — measures of **perceived** juncture between adjacent words — and should therefore be labelled upon careful consideration of the sound record. In addition, they will typically have observable physical correlates, such as tonal markings. J_ToBI currently distinguishes 4 degrees of disjuncture (on a scale from 0 (weak) to 3 (strong)) in the prosodic structure of Japanese.

All junctures (including filled pauses, cut-off words before restarts, etc.) should be assigned a break index value. The break labels should be aligned exactly with the word labels.

0 This break index marks junctures which are common in fast speech processes (e.g. /kore+wa/ → [korya] 'this-TOP').

1 This break index marks the juncture between two consecutive "words", with no higher-level prosodic boundary.

2 This break index marks a medium degree of disjuncture between adjacent words. The boundary marked by BI 2 is stronger than that marked by BI 1, but it lacks the cues (e.g. lengthening, pauses, etc.) common to an even stronger boundary marked by BI 3. In most cases the unit marked by a BI 2 at its edge corresponds to the tonally-defined unit *accentual phrase* (but see below for cases of mismatch).

3 This break index marks a strong degree of disjuncture between adjacent words, or between a word and following silent interval. BI 3 often corresponds to the boundary of the tonally-defined *intonation phrase* (but see below for cases of mismatch).

Diacritics Attached to Break Index Values

- This diacritic marks labeller uncertainty about the degree of disjuncture between adjacent words. In such cases of uncertainty, the higher level break index should be chosen, and this diacritic should be affixed directly after it (e.g. "2-" indicates uncertainty between "1" and "2").

m This diacritic marks mismatch between the subjective degree of disjuncture (break index value) and the tonal characteristics (e.g. boundary tones, downstep, etc.) In such cases of mismatch, the labeller should mark the break index according to her/his evaluation of the degree of disjuncture, and should affix this diacritic directly after it. ("2m" indicates, for example, that there is a sense of medium disjuncture that typically corresponds to an accentual phrase break, but with the tonal markings of a full intonation phrase).

p This diacritic marks a disfluent juncture marked by an abrupt cut-off or lengthening due to hesitation. This diacritic should be affixed directly after the break index value. ("1p" marks an abrupt cut-off which lacks the sense of an accentual phrase final L%, and "2p" marks a disfluent juncture accompanied by this sense of L%).

Finality Tier

This tier provides a measure of perceived finality of each intonation phrase (break index 3). At present this is a simple binary choice between "final" and "not final" (additional labels will be added once we have a more complete model of discourse finality and its relation to the structuring of information or turn-taking). A phrase which is judged as "final" will have at its edge a strong sense of disjuncture, marked by F0 lowering, segmental lengthening, creaky voice, etc. This percept of finality should be marked in this tier with the "final" label, which is aligned with the break index label. Those phrases which the labeller perceives as non-final should have no marking.

Miscellaneous Tier

This tier is reserved for other phenomena present in the speech signal which cannot be properly described by the phonological events marked in the tone and break index tiers. Such phenomena include repairs, disfluencies, silences, laughing, etc. Those phenomena which span clearly defined intervals (such as silences, laughing, etc.) should be marked by a pair of labels at their temporal beginnings and ends (e.g. laugh< .... laugh>). Other phenomena which are less clearly defined temporally should be marked by a single flag (e.g. "disfl") at the approximate location. Labeller comments may also be included in the miscellaneous tier, or additional tiers can be added to fit the needs of research at each particular site.

Appendix D Practice Utterances

Ten unlabelled utterances are included at the end of the guide for labellers to practice transcribing using the J_ToBI system. These data files are also available online in various formats on the WWW page mentioned in section 7.

Since our goal is to develop a system for labelling Japanese prosody that can be used to quickly and consistently transcribe Japanese utterances, we would like to check whether it is possible, after going through the examples and discussions in this guide, for labellers to do just this. We would therefore like to ask for your cooperation in helping us check labeller consistency of the 10 utterances included here. We would like to ask those interested to label these few utterances, and share their labels with us (either by sending us the ASCII label files, or a faxed copy of the labelled utterances). It is only by comparing the transcriptions of many people that we will be able to get an idea of how well the J_ToBI system can be used, and in which areas the difficulties lie.

Please contact the author at: `venditti@ling.ohio-state.edu` for further information. We thank you for your cooperation.

List of Practice Utterances: (in order of difficulty)

≪door≫
≪nondenai≫
≪futatsu≫
≪tugi_ni≫
≪ima_double≫
≪tree≫
≪nokotta≫
≪hachiue≫
≪migi≫
≪kasanete≫

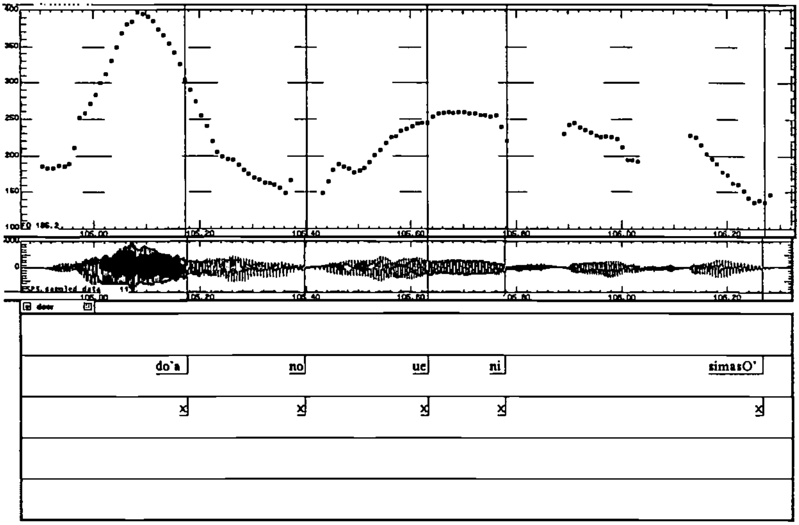


Figure 19: ≪**door**≫ "Let's put it above the door."

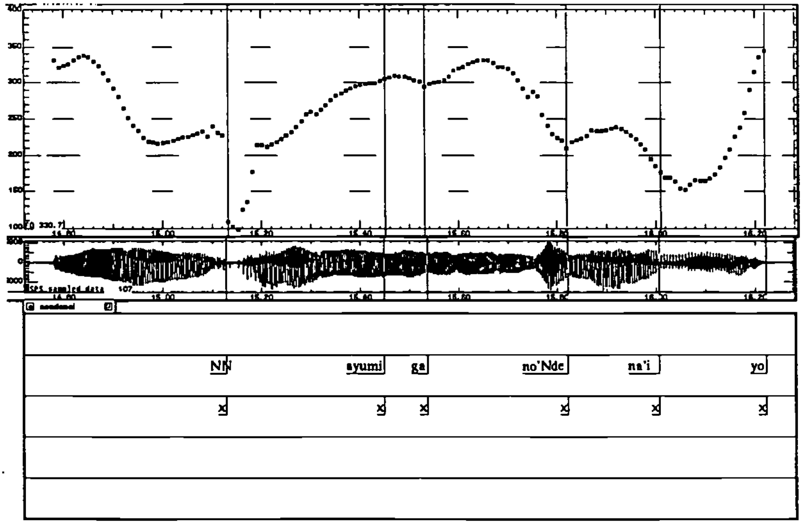


Figure 20: ≪**nondenai**≫ "No, Ayumi isn't drinking." [Don't worry about the tones on /NN/.]

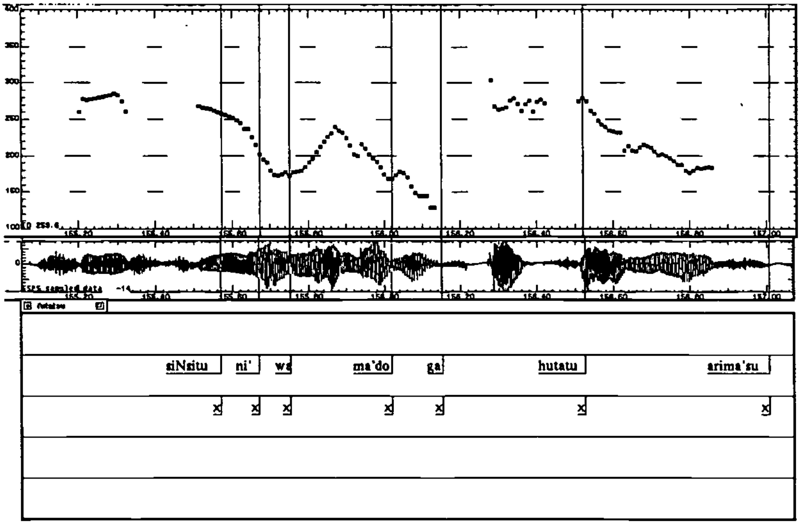


Figure 21: ≪futatsu≫ "There are two windows in the bedroom."

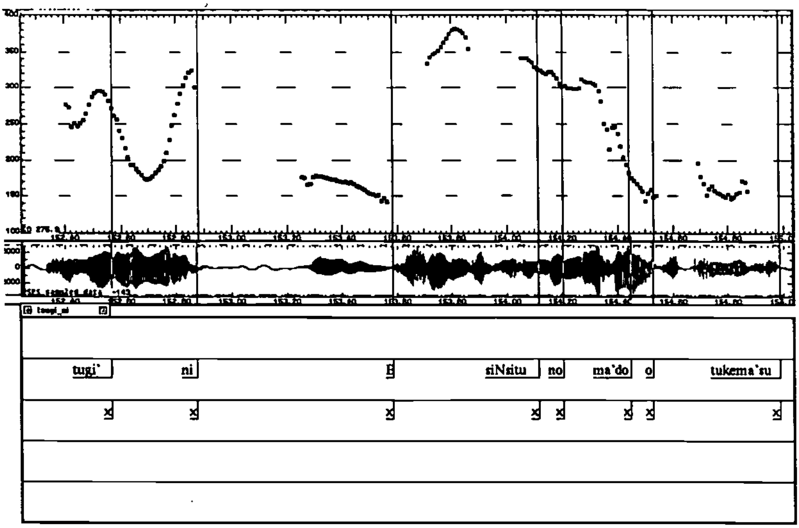


Figure 22: ≪tugi_ni≫ "Next, um, I'll attach the bedroom window."

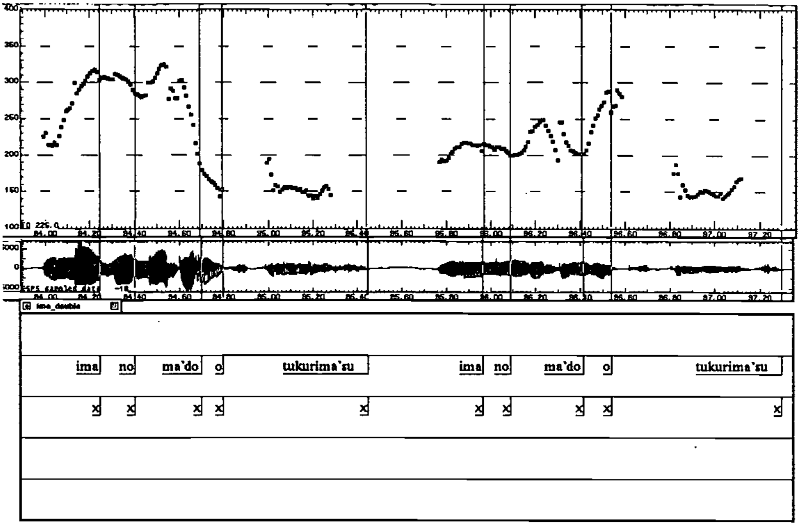


Figure 23: ≪ima_double≫ "I'll make the livingroom window. I'll make the livingroom window."

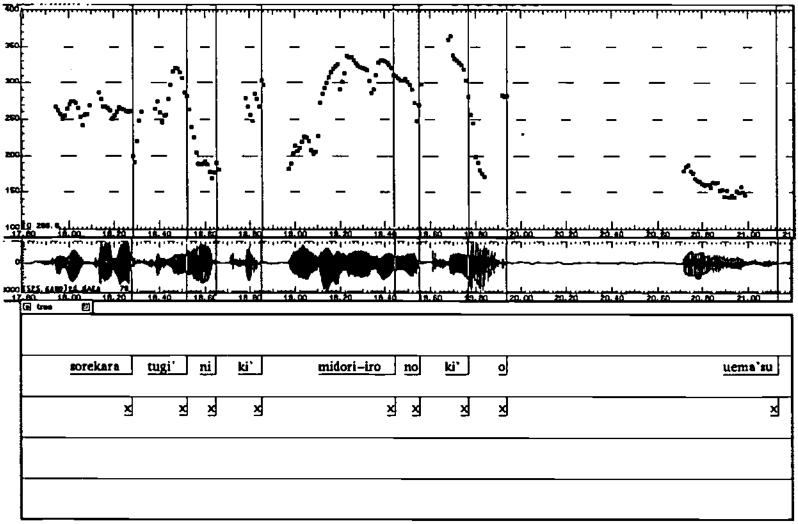


Figure 24: ≪tree≫ "Then next, I will plant a tree, a green tree."

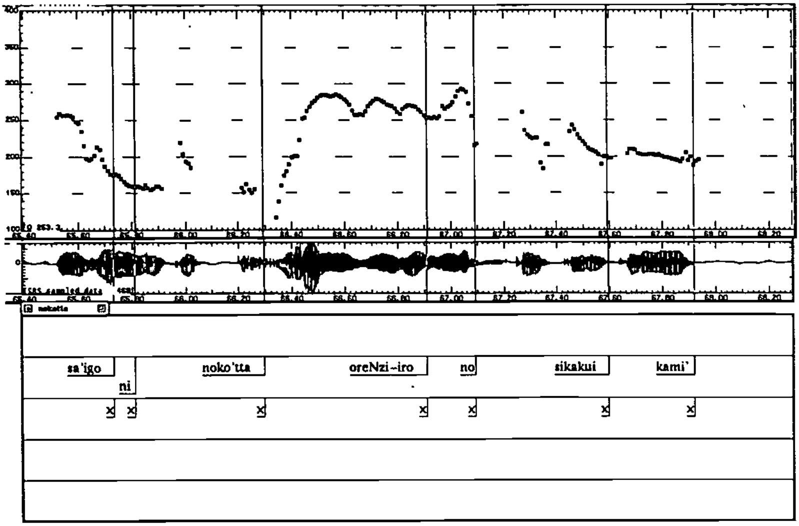


Figure 25: ≪**nokotta**≫ "the last remaining orange square paper"

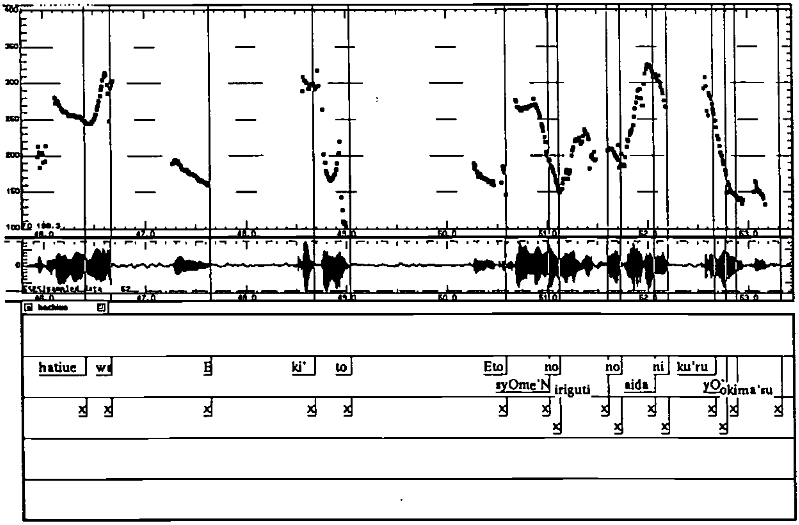


Figure 26: ≪**hachiue**≫ "I'll put the plant, um, so that it comes right in between the tree and, um, the front entrance."

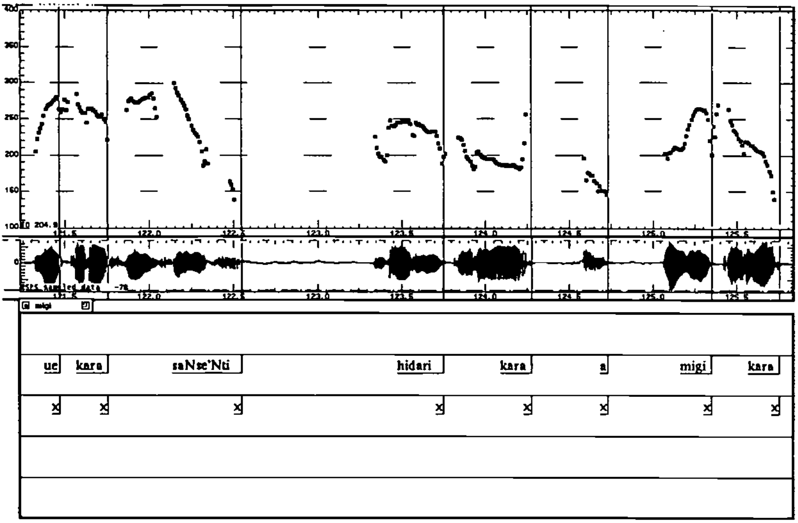


Figure 27: ≪migi≫ (Part 1) "3cm from the top, from the left ... oh, 1cm from the right, I'll attach the livingroom window here."

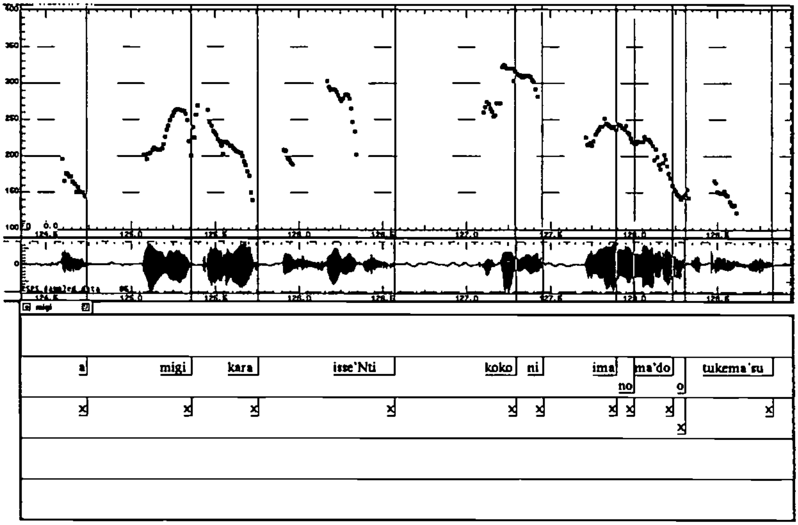


Figure 28: ≪migi≫ (Part 2) "3cm from the top, from the left ... oh, 1cm from the right, I'll attach the livingroom window here."

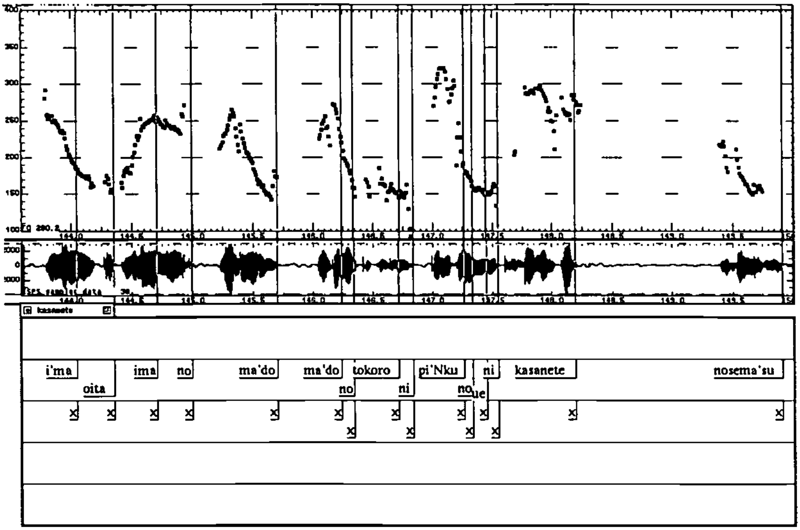


Figure 29: ≪**kasanete**≫ "I'll lay it on top of the pink one, on top of the livingroom wi- window that I just put down."

OSU Working Papers in Linguistics 50 (163-175)

A cross-linguistic study of diphthongs in spoken word processing in Japanese and English*

Kiyoko Yoneyama
yoneyama@ling.ohio-state.edu

Abstract: This paper investigates the proper treatment of diphthongs in Japanese and English in terms of spoken word processing. Three phoneme-monitoring experiments were conducted with three different groups of language users: Japanese monolinguals, English monolinguals and semi-bilingual Japanese speakers of English; both English and Japanese materials were used. The results showed that English monolinguals treat diphthongs as single units during language processing, while Japanese monolinguals treat them as two separate units. The processing of Japanese and English diphthongs by semi-bilingual Japanese speakers of English is also discussed.

INTRODUCTION

Spoken words have a rich structural organization in memory consisting of both syllabic and subsyllabic representations. Previous cross-linguistic investigation into segmenting speech has shown that word processing in different languages involves a variety of linguistic units. Results from French, for example, suggest that listeners in that language segment speech at syllable boundaries (Mehler et al, 1981). English listeners, in contrast, segment speech at the onset of strong (but not weak) syllables (Cutler and Norris, 1988). Furthermore, Japanese listeners segment speech at mora boundaries (Otake et al., 1993; Cutler and Otake 1994; Otake et al., 1996; Yoneyama, 1995). Collectively, these strategies are referred to as "rhythm-based segmentation strategies (RSS)".

In addition to an RSS, auditory recognition models such as the Cohort model (Marslen-Wilson & Welish, 1987), TRACE (McClelland & Elman, 1986), and SHORTLIST (Norris, 1994), all assume that segmentation of speech into words is achieved as an automatic consequence of lexical access from the sub-lexical phoneme- or feature-recognition process. Therefore, adult listeners with mature lexicons can map speech onto words directly without reference to language-specific prosodic units. This type of segmentation is referred to as the "general segmentation strategy (GSS)", and is available for all language users.

The focus of the present study is Japanese listeners' sensitivity to moraic structure. The Japanese mora is a component of syllable structure. Light syllables consist of one mora while syllables with a complex vowel or a coda consist of two. Japanese has a strict phonological structure and there are only five types of mora: CV, CCV, V, nasal coda (represented as N) and geminate consonant (represented as Q).

* This study is supported by Department of Linguistics and Department of Psychology. I would like to express my sincere gratitude for valuable comments and guidance provided by Mary Beckman, Jose Benki, Anne Cutler, Keith Johnson, James McQueen, and Takashi Otake. Thanks also to Mark Pitt, who enabled me to test the English listeners, and Katherine Smith and Lisa Shoaf for their technical support. I also extend my gratitude to editors, Kim Ainsworth-Darnell and Mariapaola D'Imperio for their comments on my earlier version of this paper. Of course, all errors are mine.

Japanese listeners appear to be able to exploit the moraic structure of words in a variety of different ways. Otake et al. (1993) presented Japanese listeners with sequences of natural spoken words and asked them to indicate when they detected a word beginning with a particular CV sequence, for example, /ta/. Listeners were equally fast to identify the /ta/ target in *tanishi* and in *tanshi* because both begin with the same mora /ta/ However, when the target was /tan/, Japanese listeners had difficulty detecting the target in *tanishi* because it corresponds to the whole first mora and a part of the second. Also, Cutler and Otake (1994) found that Japanese listeners detected phoneme targets which were moras in themselves more rapidly than targets which formed only part of a CV mora. Thus, for example, /n/ was detected more rapidly in *inka* than in *inori*. /o/ was detected more rapidly in *aokabi* than in *kokage*.

The vowel target /o/ in Cutler and Otake (1994) was a moraic vowel, or one that was a syllable by itself. However, a moraic vowel can occur tautosyllabically, following another vowel. Such vowel-vowel sequences are often called diphthongs. According to Vance (1987), /ai/, /ei/, /oi/, /ui/, /au/ and /ou/ can be considered diphthongs in Japanese, but only if they occur in a single morpheme and if the second segment does not bear an accent. Assuming that the previous studies concerning Japanese listeners' sensitivity to mora are correct, these people should have no problem detecting the second part of a diphthong because it is considered as one mora. Experiment 1 tests this hypothesis. Hypothesis is a comparison: If Japanese listeners process speech mora by mora in a mora-based RSS, then they should identify [i] in a diphthong (a V mora) more quickly and accurately than [i] in a CV mora.

EXPERIMENT 1

PARTICIPANTS

Twenty-four Dokkyo University undergraduates participated in exchange for course credit. All were native speakers of Japanese and reported no hearing difficulties. All had studied English only in school and none had ever stayed in an English-speaking country for more than 3 months.

MATERIALS AND PROCEDURE

Two sets of language materials (Japanese and English) were constructed. The Japanese materials, shown in Table 1, consist of 24 content words (nouns, verbs, and adjectives). Half contained [oi] (CV_1V_2 words) and the other half contained [i] (CV_1 words). They formed twelve pairs, contrasting in the occurrence of [o] in the first syllable. Each of CV_1V_2 words contained a diphthong in Japanese, which follows Vance's (1987) definitions. We admit that because a number of words which contain an [oi] sequence is limited in Japanese, most of the stimuli are borrowed words[1].

CV_1V_2 words			CV_1 words		
koin	boisu	join	rin	nisu	misu
boiru	doitsu	roido	misa	piza	bika
koiru	hoiru	hoiro	pin	kika	piru
noizu	koika	toire	bisu	biru	kigo

Table 1: Japanese stimulus words

[1] Of CV1V2 words, only "koika" is a Sino-Japanese word, and the others are borrowed words.

The English materials, shown in Table 2, are very similar in the Japanese materials. As before, there are 24 content words (nouns, verbs, and adjectives). Half contained [oɪ] (CV_1V_2 words) and the other half contained [ɪ] (CV_1 words). They formed twelve pairs, contrasting in the occurrence of [o] in the first syllable.

CV_1V_2 words			CV_1 words		
join	boil	moist	pit	kin	mist
foil	coil	soil	gin	bill	win
toil	voice	loin	fill	kill	sit
coin	void	joint	till	mill	fit

Table 2: English stimulus words

The vowel sequences [oi] and [oɪ] were chosen for my stimuli because they are represented by two letters in both English and Japanese. Thus, any orthographic influence on responses should be equal across languages.

The English materials were recorded by a male native speaker of American English (Northern Kentucky dialect), while the Japanese materials were recorded by a male native speaker of Japanese (Tokyo dialect). All materials were stored on DAT tapes and spoken at a normal rate of speech.

The target words were mixed with filler words and were arranged into 48 word sequences in both sets of materials. In each set of materials, half contained one of the experimental target words and the other half did not. Each word sequence varied from two to six words. Of the twenty-four sequences which did not contain one of the experimental target words, the half contained a dummy target, which elicited participants' responses. A target always occurred in the penultimate position in each word sequence.

Stimuli were presented to participants binaurally over headphones at a comfortable listening level. Responses were collected by pressing a button on a response board. Stimulus presentation and data collection were controlled by a PC computer. Participants were tested in a quiet room. The English and Japanese materials were presented in separate session, with the English session coming first. In each test session, participants listened to 48 word sequences. For each sequence, they were instructed to think constantly of the target sound and to respond as soon as they detected it. The target sound in each test session was different. In the English test session, it was a high front lax vowel ([ɪ]) and in the Japanese test session, it was a high front tense vowel ([i]). Participants were not informed about the experimental manipulations or about the characteristics of the stimuli. A 10-sequence practice session preceded the 48 test sequences, which were presented in a fixed order in each test session. The entire experiment lasted approximately 50 minutes.

RESULTS

Analyses of variance were conducted separately on the mean rates of missed data and mean response times. Miss rate is a percentage of trials on which the listeners failed to detect the target. RT rate is time in ms between the target sound onset and the listener's detection response. The condition means of miss rate are give in Figure 1. The analysis of miss rate showed that Japanese targets were detected significantly more accurately (19.3%) than English ones (87.6%; F [1, 1148] = 1043.439, p< .0001). Also, targets in CV_1V_2 words were detected more accurately (50%) than those in CV_1 words (57.2%; F [1, 1148] = 11.242, p< .001). There was an interaction between material language and word type: targets in CV_1 words in the Japanese materials were detected more accurately than those in the English materials (F [1, 1148] = 4.875, p < .05). In English materials, targets in CV_1V_2 words were detected more accurately than those in CV_1 words (F [1, 576] = 19.366, p < .0001). However, the same effect was not observed in the Japanese materials.

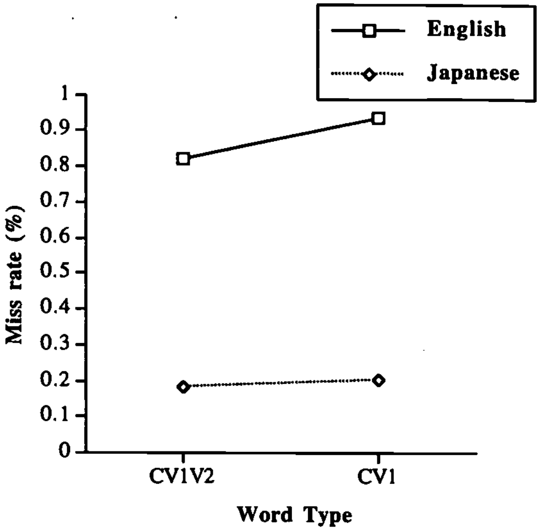


Fig.1: The mean miss rate in (%) as a function of stimulus word type and material language, Experiment 1.

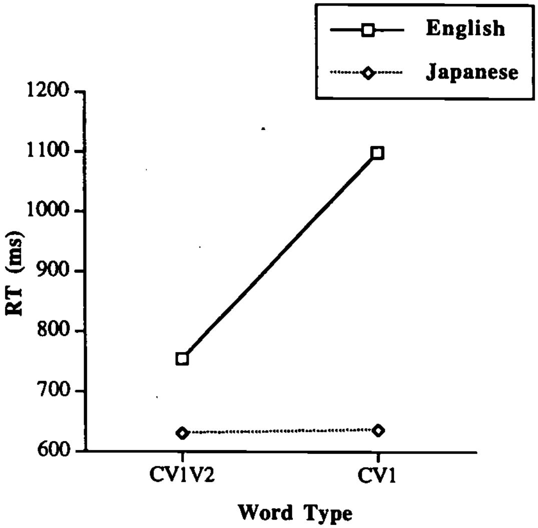


Fig.2: The mean response times in ms. as a function of stimulus word type and material language, Experiment 1.

The condition means of response times are give in Figure 2. The analysis of response times showed that Japanese targets were detected significantly faster (633 ms) than English ones (844 ms; F [1, 531] = 1.07, p< .0001). It also showed that targets in CV_1V_2 words were detected significantly faster (654 ms) than those in CV_1 words (668 ms; F [1, 531] = 37.11, p< .0001). There was an interaction between material language and word type: in the English materials, targets in CV1V2 words were detected significantly faster (756 ms) than those in CV_1 words (307 ms), and in CV_1 words, Japanese targets were detected significantly faster (631 ms) than English ones (1100 ms; F [1, 531] = 35.565, p< .0001).

DISCUSSION

Experiment 1 provides contrasting evidence concerning whether or not Japanese listeners use general and rhythm-based segmentation strategies. There was no evidence of a moraic effect with the Japanese materials: responses to the targets in CV_1V_2 words were as fast and accurate as those in CV_1 words. If the participants had employed moraic segmentation, they should have been able to detect targets in CV_1V_2 words significantly more quickly and accurately than those in CV_1 words. The responses to English materials, conversely, suggest that Japanese listeners are sensitive to the moraic structure, which is consistent with Cutler and Otake (1994). Both the miss rate analysis and the response time analysis showed a moraic effect: targets in CV_1V_2 words were detected significantly faster and more accurately than those in CV_1 words.

One possible explanation for these results is that the participants employed a GSS when they listening to Japanese and an RSS when listening to English. Cutler et al. (1992) have hypothesized that two segmentation strategies are available for processing of a native language. Young children use an RSS to construct their lexicon. As their lexicon matures, however, they learn to segment speech directly into words, without any phonological units intervening (GSS). Since all participants in this experiment have a mature Japanese lexicon, it is not surprising that they may have employed a GSS in the processing of Japanese, their native language. This would be consistent with the findings in Yoneyama (1995), where Japanese listeners showed the same GSS as English and French listeners in Cutler et al. (1986).

My explanation for the responses to English materials is consistent with Cutler and Otake (1994). Cutler and Otake reported that their Japanese participants are sensitive to moraic nasals even in English, as if they were listening to Japanese. This could indicate that Japanese listeners employ an RSS when they are listening to a non-native language. My Japanese participants showed moraic sensitivity to the vowels in English diphthongs.

One concern regarding my conclusion here is why the Japanese participants did not show their sensitivity to moraic structure in Japanese diphthongs, even though all previous on-line experiments have found such effect. One possible explanation is that Japanese listeners are also sensitive to syllable structure. Notice that Cutler and Otake (1994) showed a clear moraic effect using vowel targets in Japanese materials. On the other hand, the current study did not show this effect even though targets were second half of the vowel sequence. The only difference between two studies was position of targets in terms of syllable structure. Each of Japanese vowel targets of Cutler and Otake (1994) was also considered as one syllable by itself. Conversely, each of my Japanese vowel targets was a part of nucleus and it did not become a syllable by itself. Another possibility is that Japanese listeners treat diphthongs as single units even when they listening to Japanese. Thus, since the bond between two vowels in diphthongs is so tight, the second vowel of a diphthong was hard to be detected even though it had a moraic status. In any case, this inconsistency is certainly a topic for future study.

In Experiment 2, I will examine how English listeners treat diphthongs in English and Japanese. Both Otake et al. (1993) and Cutler and Otake (1994) found that English speakers did not show any sensitivity to moraic structure. Cutler et al. (1992) hypothesized that the English participants in these studies employed an RSS when listening to Japanese. If this is true, English listeners in the current study should also use the same listening strategy with English and Japanese materials, and show no sensitivity to moraic structure.

EXPERIMENT 2

PARTICIPANTS

Nineteen Ohio State University undergraduates participated in exchange for course credit. All were native speakers of English and reported no hearing difficulties. None have ever either studied Japanese or stayed in Japan for more than 3 months.

MATERIALS AND PROCEDURE

The materials were the same as in Experiment 1. However, in this experiment, all the English and Japanese stimulus words were digitally recorded onto computer disk (sampling rate of 10 kHz, low-pass filtered at 4.8 kHz) and edited and saved as separated sound files. Stimuli were presented to participants binaurally over headphones at a comfortable listening level. Responses were collected by pressing the button on a response board. Stimulus presentation and data collection were controlled by a PC/AT computer.

The procedure was generally the same as in Experiment 1, with a few minor changes. As many as four participants at a time were tested simultaneously in individual sound-attenuated booths. Also, the order of two test sessions was counterbalanced; half of the participants were presented the English test session first and the latter half were presented the Japanese test session first. The entire experiment lasted approximately 50 minutes.

RESULTS

Analyses of variance were conducted separately on the mean rates of missed data in each condition and the mean response time in each condition, separately. The condition means of miss rate are give in Figure 3. The analysis of miss rate showed that the target in CV_1 words was detected more accurately (36.7%) than that in CV_1V_2 words (72.2%; F [1, 908] = 133.281, p< .0001). The results showed an interaction between material language and word type: the target in CV_1 words in the Japanese materials was detected less accurately than that in the English materials and the target in CV_1V_2 words in the English materials was detected less accurately than that in the Japanese materials (F [1, 908] = 6.582, p< .0105). No effect of material language was observed. Further analyses showed that in both language materials, CV_1 words were detected more accurately than CV_1V_2 words (F [1, 454] = 106.599, p < .0001 for English; F [1, 454] = 37.813, p < .0001 for Japanese).

The condition means of response times are give in Figure 4. The analysis of response times showed that Japanese stimulus words were detected significantly faster (715 ms) than English ones (788 ms; F [1, 386] = 25.17, p< .0001). In addition, targets in CV_1 words were detected significantly faster (725 ms) than that CV_1V_2 words (806 ms; F [1, 386] = 19.996, p< .0001). There was an interaction between material language and word type: in both languages, targets in CV_1 words were detected significantly faster than in CV_1V_2 words (F [1, 386] = 13.301, p < .0003). However, the response time difference between two conditions in English was much larger (206 ms) than that in Japanese materials (21 ms).

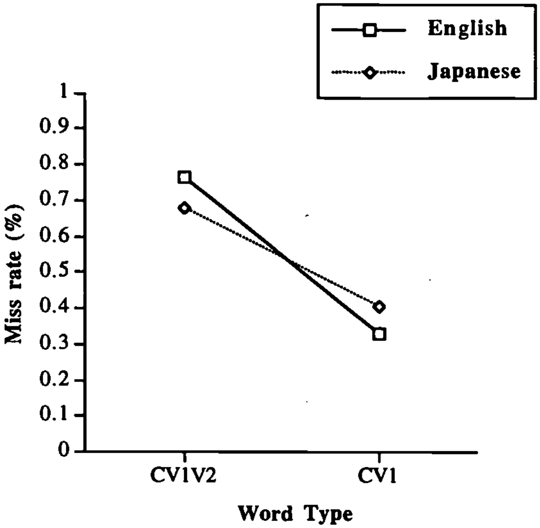


Fig.3: The mean miss rate in (%) as a function of stimulus word type and material language, Experiment 2.

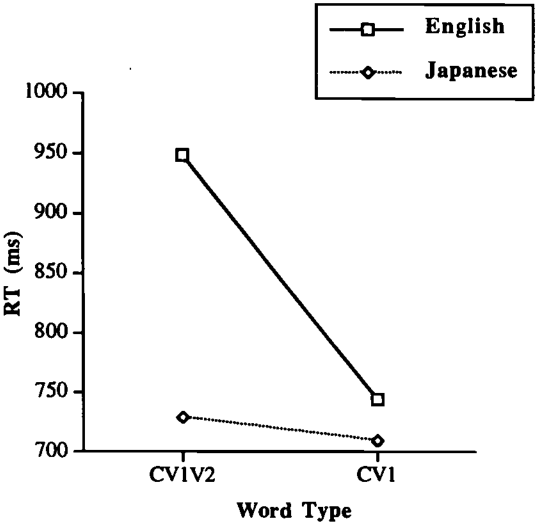


Fig.4: The mean response times in ms. as a function of stimulus word type and material language, Experiment 2.

DISCUSSION

The miss rate analysis showed that English listeners treated diphthongs in English and Japanese in the same way. They consistently responded to targets in CV_1 words more accurately than to CV_1V_2 words. The analysis of response time showed the same.effect. If they were sensitive to moraic structure, the results should be opposite: targets in CV_1V_2 words were detected faster and more accurately than those in CV_1 words. This not being the case, it appears that my English participants were not sensitive to moraic structure.

My results here confirmed the previous on-line studies with English listeners when listening to Japanese. Both Otake et al. (1993) and Cutler and Otake (1994) reported that English speakers seemed to process Japanese as if they were listening to English. It seems feasible that the English listeners in my study may have employed a native way of listening even when listening to a non-native language. In other words, my English participants might have employed their English RSS when listening to Japanese.

One concern regarding this conclusion is that my experiment here cannot directly show the English speakers' stress-timed segmentation strategy, which was originally confirmed by Cutler and Norris (1988). These researchers found that English listeners segment speech at the strong syllables, which needs the vertical specification of speech (strong versus weak stress assignment). However, I assume that the effect of material language in reaction time analysis might show some language-specific way of listening by my English participants.

Notice that my English participants listened to two set of language materials, both English and Japanese language materials, and that in each set, the target sound was different. In the English materials, the target sound was always [ɪ] whereas in the Japanese materials, it was always [i]. Thus, the effect of material language might be rephrased as the effect of the target sound difference in two languages. The reaction time analysis might indicate that [i] can be detected faster than [ɪ] across languages. Findings in van Ooijen (1994) support this idea. She investigated the processing of vowels and consonants by English listeners, and conducted 9 reaction-time experiments. Using her list of individual target bearing items together with their mean reaction time in (ms) across participants, it is possible to estimate the processing time of these two vowels. Table 3 shows the from van Ooijen's (1994) Experiments 5 and 9, that have the same structure of my English materials. All the words have the primary stress on the first syllable. An analysis of variance I calculated from these words showed that */i/* was detected significantly faster (414 ms) than */ɪ/* (531.5 ms; F [1, 12] = 23.06, p< .004), which is consistent with my finding.

words with /i/		words with /ɪ/	
feast	458	tissue	494
sheep	483	mitten	581
priest	439	liver	514
seek	354	wither	502
leave	358	sickle	525
scream	397	mistress	573
needle	373		
feeling	454		
Mean	**414.5**	**Mean**	**531.5**

Table 3: words partially from Experiments 5 and 9 in van Ooijen (1994) that contain either */ɪ/* or */i/*.

In any case, we can be sure that acoustic information can affect participants' performance when processing spoken materials in one's native and non-native language. This suggests that understanding the mechanism between the segmentation and acoustic information is crucial for revealing spoken word processing.

Experiment 3 investigates how semi-bilingual Japanese speakers of English treat English and Japanese diphthongs in spoken word processing. Of interest is whether or not knowledge of a second language can influence processing in the first, or vice versa. Participants in this experiment were native speakers of Japanese who has lived in an English speaking country for more than 3 years (semi-bilinguals). Because Japanese is their first language, I expect that they will treat Japanese diphthongs like their monolingual Japanese counterparts. But how will these individuals treat English diphthongs? Cutler et al. (1992) have hypothesized that bilinguals and monolinguals perform differently when they listen to a non-native language. This might indicate that diphthongs in English words stored in the lexicon of semi-bilinguals might be treated differently from those for Japanese words, depending on the degrees of foreign-language exposure. The last experiment explores this possibility.

EXPERIMENT 3

PARTICIPANTS

Twenty-two Dokkyo University undergraduates participated in exchange for a nominal fee. All had good communicative abilities in English, and had lived for a minimum of three years in an English-speaking country, but were recognizably not native speakers of English. They reported no hearing difficulties.

MATERIALS AND PROCEDURE

Materials and procedure were the same as in Experiment 1.

RESULTS

Analyses of variance were conducted separately on the mean rates of missed data in each condition and the mean response time in each condition. The condition means of miss rate are give in Figure 5. The analysis of miss rate showed an interaction between material language and word type: targets in CV_1V_2 words were detected significantly more accurately (14%) than those in CV_1 words (18.9%) in Japanese, and CV_1 words were detected significantly more accurately (12.1%) than CV_1V_2 words in English (19.3%; F [1, 1052] = 7.204, p< .01). No other main effect was observed.

The condition means of response times are give in Fig. 6. No main effect or an interaction were observed.

DISCUSSION

This experiment with semi-bilinguals confirmed the previous findings with Japanese monolingual and bilingual listeners. First, the analysis of miss for Japanese materials showed that my semi-bilingual listeners are sensitive to moraic structure when listening to Japanese. They responded to targets in CV_1V_2 words more accurately than to those in CV_1 words, suggesting that they employed a Japanese RSS. This finding is consistent with previous on-line studies with Japanese listeners (Otake et al., 1993; Cutler et al. (1994). Another finding from the miss rate analysis is that my semi-bilinguals might have employed a GSS when listening to English, which is consistent with the findings by semi-bilinguals Japanese speakers of English in Experiment 1 in Yoneyama (1996).

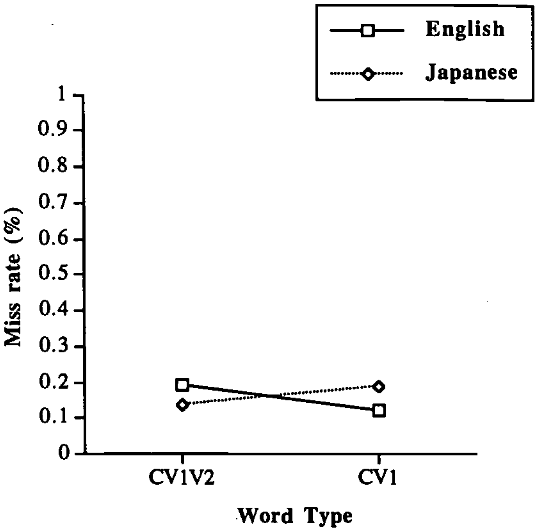


Fig.5: The mean miss rate in (%) as a function of stimulus word type and material language, Experiment 3

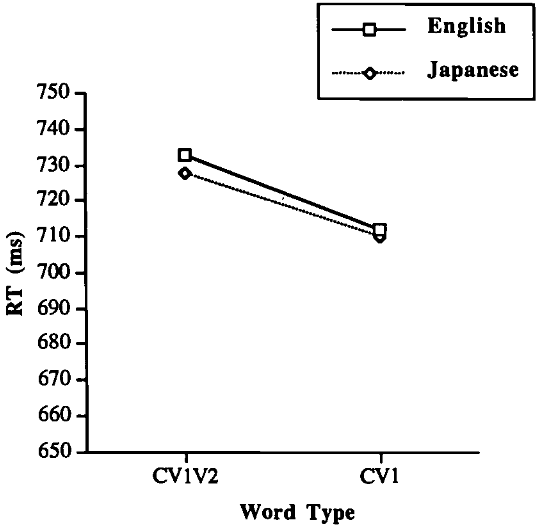


Fig.6: The mean response times in ms. as a function of stimulus word type and material language, Experiment 3.

These findings might indicate that semi-bilinguals can choose either a GSS and an RSS, depending on the language input. One possible explanation is that the second language exposure might enable them to shift the application of their RSS (moraic segmentation) which do not work well with English, to the one of a GSS. This is exactly what Cutler et al. (1992) have claimed: bilingual speakers have an ability to suppress their native segmentation strategy to a non-native language.

Of course, it is possible that the semi-bilinguals in the current study may have been sufficiently fluent in both the first and second languages to rely on segmental information no matter which segmentation strategy they are using. The miss rate in the four conditions were very low and the overall miss rate was 15.5%. Surprisingly, even when we compare them with English listeners in CV_1 word condition in English materials, performance of the semi-bilinguals was better (12.1%) than that of English listeners (32.9%) from Experiment 2. Also, the results of response time analysis showed neither significant main effects nor an interaction. These findings might suggest that the semi-bilingual participants may be able to use segmental information equally in English and in Japanese.

GENERAL DISCUSSION

The current study further investigates the sensitivity to moraic structure of three groups of language users: Japanese listeners, English listeners and semi-bilinguals Japanese speakers of English. Experiment 1 was designed to further investigate the sensitivity to moraic structure by Japanese listeners, using two sets of language materials (English and Japanese). The Japanese listeners showed a moraic effect in English materials, although they did not show it in Japanese materials, where it has been found in the previous studies. Experiment 2 investigated how English listeners responded to the same materials used in Experiment 1. The results showed that the English listeners were not sensitive to moraic structure either in English or Japanese, and seemed to listen to both English and Japanese in the same way. Experiment 3 tested how semi-bilingual Japanese speakers of English responded to the same materials used in Experiment 1. The results indicated that the semi-bilinguals showed a moraic effect in Japanese materials while they did not in English materials.

The possible comparisons among these three different groups of language users shed light on further aspects of segmenting speech. First of all, from the results in Experiments 1 and 3, we see native speakers of Japanese are generally sensitive to moraic structure. The semi-bilinguals showed their sensitivity in Japanese materials, which clearly supports Otake et al. (1993) and Cutler and Otake (1994).

One concern with our conclusion is that my Japanese listeners in Experiment 1 did not show a clear moraic effect in Japanese materials, which was supposed to show from the findings in the previous studies. Even though we have not yet figured out why my Japanese listeners employed a unexpected GSS to process Japanese diphthongs, I assume that some unknown factor, such as syllable structure and acoustic property of diphthongs affected my participants' performance, and made them switch their listening strategy from an RSS to a GSS. Since my Japanese listeners showed their sensitivity even to English materials, which indicates a typical characteristic of responses by monolingual listeners, I hypothesize that the Japanese participants in Experiment 1 are sensitive to moraic structure, even though the results did not show it as clearly as was hoped. Both Experiments 1 and 3 showed the sensitivity to moraic structure by native speakers of Japanese. They confirmed the previous on-line studies with Japanese listeners.

A comparison between native and non-native speakers of Japanese will allow us to explore the language-specific aspects of listening. As we have shown above, native speakers of Japanese showed the sensitivity to moraic structure. Conversely, Experiment 2 showed, that English listeners did not show the same sensitivity, as in the previous studies. These results confirmed that only Japanese natives are sensitive to moraic structure.

In addition to the sensitivity to moraic structure by Japanese natives, the data of semi-bilinguals in this study suggests that extensive second-language experience enables people to treat diphthongs differently in different languages. As Cutler et al. (1992) have proposed, the main difference between bilingual and monolingual speakers is that the former have a lexicon of a familiar non-native language whereas the latter does not. This means that semi-bilingual Japanese speakers must have acquired English words in their lexicon as well, even though English is their second-language. Their extensive language experience enables bilingual speakers to suppress their native rhythm-based strategy when dealing with foreign language input. Experiment 3 showed that semi-bilinguals treat diphthongs differently, depending on which language they are listening to.

Finally, I would like to mention about the relation between diphthongs and language processing. Japanese listeners in the current study did not have difficulty detecting the second vowel of diphthongs in English and Japanese. This suggests that they treat diphthongs as two separate units. On the other hand, English listeners in Experiment 2 showed a difficulty detecting the same targets in the same materials. This might indicate that they treat both Japanese and English diphthongs as single units in the language processing. Interestingly, the semi-bilinguals in Experiment 3 seem to treat English and Japanese diphthongs differently. In Japanese, they seem to employ a moraic segmentation, and to treat Japanese diphthongs as two separate units, just like monolingual Japanese listeners in Experiment 1. On the other hand, when they listen to English, they seem to treat English diphthongs differently. In English materials, they responded to targets in CV_1 word more accurately than to those in CV_1V_2 words. This might indicate that they employ a general segmentation strategy. Also a clear interaction between language material and a word type might suggest that they also treated English diphthongs as single units, like English monolinguals in Experiment 2. However, as we have seen in Experiments 2 and 3, the semi-bilinguals listening to English detected a high front lax vowel in CV_1V_2 words in 80% of the trials, even though English listeners detected the same target in the same materials in only 20% of the trials. They were definitely able to identify the target sound in English diphthongs whereas English monolinguals were not. This might also suggest that performance by semi-bilinguals when processing in English is different from that by English monolinguals. Together, the results in Experiment 3 might be interpreted as follows. Firstly, second language experience may enable individuals to treat diphthongs differently in that language. This language knowledge enabled them to switch from an RSS to a GSS, which confirmed the previous bilingual studies (Cutler et al., 1992; Bradley et al., 1993; Kearns 1994; Yoneyama 1996). Secondly, however, they cannot fully suppress their native way of listening. The results showed that they did not have difficulty accessing to moraic structure even in English. This is because they are not able to fully suppress their native RSS when they processing in English. My semi-bilinguals are not "perfect" bilinguals as those in Cutler et al. (1992). This finding is coincide with the findings with semi-bilingual Japanese speakers of English in Yoneyama (1996).

In this paper, I claim that, in terms of spoken language processing, Japanese diphthongs should be treated as two separate units in reference to moraic structure, and English diphthongs should be treated as single units. However, I do not attempt to represent Japanese and English diphthongs in syllable structures here. This should be investigated in further studies.

REFERENCES

Bradley, D.C., Sanchez-Casas, R.M., & Garcia-Albea, J.E. (1993). "The status of the syllable in the perception of English and Spanish," *Language and Cognitive Processes, 8, 197-233.*

Cutler, A., Mehler, J., Norris, D. G. & Segui, J. (1992). "The monolingual nature of speech segmentation by bilingual," *Cognitive Psychology*, 24, 381-410.

Cutler, A. & Norris, D. G. (1988). "The role of strong syllables in segmentation for lexical access," *Journal of Experimental Psychology: Human Perception & Performance*, 14, 113-121.

Cutler, A. & Otake, T. (1994). "Mora or Phoneme? Further evidence for language-specific listening," *Journal of Memory & Language*, 33, 824-844.

Kearns, R.K. (1994). *Prelexical speech processing in mono- & bilinguals*, Ph.D. Dissertation, University of Cambridge.

Marslen-Wilson, W.D. & Welsh, A. (1978). "Processing interaction and lexical access during word recognition in continuos speech," *Cognitive Psychology*, 10, 29-63.

McClelland, J.L. & Elman, J.F. (1986). "The TRACE model of speech perception," *Cognitive Psychology*, 18, 1-86.

Mehler, J., Dommergues, J.-Y., Frauenfelder, U. & Segui, J. (1981). "The syllable's role in speech segmentation," *Journal of Verbal Learning & Verbal Behavior*, 20. 298-305.

Norris, D.G. (1994). "SHORTLIST: A hybrid connectionist model of continuous speech recognition," *Cognition*, 52, 189-234.

Otake, T., Hatano, G., Cutler, A. & Mehler, J. (1993). "Mora or syllable? Speech segmentation in Japanese," *Journal of Memory & Language*, 32, 258-278.

Otake, T., Hatano, G., & Yoneyama, K. (1996). "Japanese speech segmentation by Japanese listeners," in T. Otake & A. Cutler eds., *Phonological structure and language processing: Cross-linguistic studies*, 183-201, Mouton de Gruyter, Berlin.

van Ooijen, B. (1994). *The processing of vowel and consonant*, Ph.D. dissertation, University of Leiden, The Netherlands.

Vance, T. J. (1987). *An Introduction to Japanese Phonology*. State University of New York Press.

Yoneyama, K. (1995). *Segmentation procedure by semi-bilingual speakers of Japanese and English*. An unpublished M.A. thesis, Dokkyo University, Soka, Japan.

Yoneyama, K. (1996). "Segmentation strategies for spoken language recognition: Evidence from Semi-bilingual Japanese speakers of English," *Proceedings of the 1996 International Conference on Spoken Language Processing*, Philadelphia, vol. 1, 454-457.