The Ohio State University

Working Papers in Linguistics No. 54

VARIA

—

Edited by
Jennifer S. Muller
Tsan Huang
Craige Roberts

The Ohio State University
Department of Linguistics

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

Autumn 2000

Available Titles of the Ohio State University
WORKING PAPERS IN LINGUISTICS
(OSU WPL)

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 , 23and 32 are out of print and no longer available. The table of contents can also be viewed on our WWW (World Wide Web) server:

http://ling.ohio-state.edu/department/osuwpl.html

There are two ways to subscribe to WPL. 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.

#21, $5.00. 252 pp. (May 1976): Edited by Arnold Zwicky: *Papers on Nonphonology*. Papers by Steven Boer and William Lycan, Marian Johnson, Robert Kantor, Patricia Lee, 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 Zwicky: *Papers in Phonology*. Papers by Donald Churma, Roderick Goman (OSU Ph.D. Dissertation), and Lawerence 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): Lawerence Clifford Schourup: *Common Discourse Particles in English Conversation*. (OSU Ph.D. Dissertation)

#29, $5.00. 207 pp. (May 1984): Edited by Arnold Zwicky & Rex Wallace: *Papers on Morphology*. Papers by Belinda Brodie, Donald Churma, Erhard Hinrichs, Brian Joseph, Joel Nevis, Anne Stewart, 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 Hinrichs, Michael Geis and Arnold Zwicky.

#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 Riitta 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 Zwicky: *A Festschrift for Ilse Lehiste*. Papers by colleagues of Ilse Lehiste at 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 Riitta 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, Yong-kyoon 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, Ken 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 Ohio State University May 26 - 27, 1990.

#40, $15.00. 438 pp. (July, 1992): Edited by Chris Barker and David Dowty: *SALT II: Proceedings from the Second Conference on Semantics and Linguistic Theory*. Contains twenty papers presented at the conference held at Ohio State University May 1 - 3, 1992.

#41, $12.00. 148 pp. (December 1992): Edited by Elizabeth Hume. *Papers in Phonology*. Papers by Benjamin Ao, Elizabeth Hume, Nasiombe Mutonyi, David Odden, Frederick Parkinson, and R. 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, and Jae-Hak Yoon.

#43, $12.00. 130 pp. (January 1994): Edited by Sook-hyang Lee and Sun-Ah Jun: *Papers from the Linguistics Laboratory*. Papers by Ken deJong, Sun-Ah Jun, Gina Lee, Janet Fletcher & Eric Vatikiotis-Bateson, Benjamin Ao, Monica Crabtree & Claudia Kurz, Sook-hyang Lee, Ho-hsien Pan, and Sun-Ah Jun & Islay Cowie.

#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, $15.00 169 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, $12.00 128 pp (October 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: Mike Cahill: Rebecca Herman, Hyeon-Seok Kang, Nasiombe Mutonyi, David Odden, Frederick Parkinson, Robert Poletto, R.Ruth Roberts-Kohno.

#49, $15.00 177pp (Spring 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, Jae-Hak Yoon.

#50, $15.00 175pp (Spring 1997): Edited by Kim Ainsworth-Darnell and Mariapaola D'Imperio: *Papers from the Linguistics Laboratory.* Papers by Michael Cahill, E. Diehm & K. Johnson, Mariapaola D' Imperio, Steve Hartman Keiser et al., Rebecca Herman, Keith Johnson, Mary Beckman & Keith Johnson, Jennifer Venditti, Kiyoko Yoneyama.

#51, $15.00 214pp (Summer 1998): Edited by Mary M. Bradshaw, David Odden and Derek Wyckoff. *Varia.* Papers by Mary M. Bradshaw, Michael Cahill, Gwang-Yoon Goh, Robert Poletto, Shravan Vasishth, and Neal Whitman.

#52, $15.00 288pp (Summer 1999): Edited by Brian Joseph. Papers by Amalia Arvanti & Brian D. Joseph, Michael Cahill, Gwang-Yoon Goh, Karin Golde, Craige Hilts, Martin Jansche, Brian D. Joseph & Catherine Karnitis, Steve Hartman Keiser, Panayiotis Pappas, Michelle F. Ramos-Pellicia, Charlotte Christ Schaengold, Thomas W. Stewart, Jr, Pauline Welby & Neal Whitman, James Weller, Marika Whaley.

53, $15.00 198pp (Summer 2000): Edited by Amanda Miller-Ockhuizen, Robert Levine, and Anthony J. Gonsalves. Papers by Alison R. Blodgett, Michael Cahill, Hope Dawson, Panayiotis Pappas, Shravan Vasishth, Neal Whitman, Steve Winters.

Please include payment with your order. We are able to process personal checks, travelers' checks, money orders drawn to an US-bank, or cash (unfortunately, we are unable to accept credit cards as a form of payment). Please make checks payable to the *Ohio State University*.

Please send your correspondence to:

OSU WPL,
Dept. of Linguistics, lingadm@ling.ohio-state.edu
Ohio State University,
222 Oxley Hall,
1712 Neil Ave.,
Columbus, OH 43210-1298

I / We would like to subscribe to the Ohio State University *Working Papers in Linguistics* on an:

a. ___ Issue by issue basis
b. ___ Regular subscription basis
c. ___ No Subscription, but send the following issues:

NAME: _____

ADDRESS: _____

_____

_____

E-MAIL: _____

Information Concerning OSDL

(Ohio State Dissertations in Linguistics)

Ohio State Linguistics Students are now making available dissertations written since 1992 by students in the linguistics department. For more information regarding available titles and abstracts as of Summer 2000, please visit our website at :

http://ling.ohio-state.edu/department/dissertations.html

For more information and ordering procedures, please contact:

OSDL or osdl@ling.ohio-state.edu
Department of Linguistics
Ohio State University
222 Oxley Hall
1712 Neil Avenue
Columbus, OH 43210-1289
USA

Currently available titles (Summer 2000):

Ainsworth-Darnell, Kim (1998) *The Effects of Priming on recognition Latencies to Familiar and Unfamiliar Orthographic Forms of Japanese Words.*

Ao, Benjamin Xiaoping (1993) *Phonetics and Phonology of Nantong Chinese.*

Arrieta, Kutz M.C. (1998) *Nominalizations in Basque: A Case In Language Attrition.*

Ayers, Gayle (1996) *Nuclear Accent Types and Prominence: Some Psycholinguistic Experiments.*

Bradshaw, Mary (1999) *A Crosslinguistic Study of Consonant-Tone Interaction.*

Cahill, Michael Clark (1999) *Aspects of the Morphology and Phonology of Konni.*

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

Chung, Chan (1995) *A Lexical Approach to Word Order Variation in Korean.*

Dai, John Xiang-ling (1992) *Chinese Morphology and its Interface with the Syntax.*

Golde, Karin E. (1999) *Binding Theory and Beyond: an Investigation into the English Pronominal System.*

Herman, Rebecca (1998) *Intonation and Discourse Structure in English: Phonological and Phonetic Markers of Local and Global Discourse Structure.*

Jun, Sun-Ah (1993) *The Phonetics and Phonology of Korean Prosody.*

Kang, Hyeon-Seok (1996) *Phonological Variation in Glides and Diphthongs of Seoul Korean: Its Synchrony and Diachrony.*

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

Kim, No-Ju (1996) *Tone, Segments, and their Interations in North Kyungsang Korean: A Correspondence Theoretic Account.*

Lee, Gina (1993) *Comparative, Diachronic and Experimental Perspectives on the Interaction Between Tone and the Vowel in Standard Cantonese.*

Lee, Sook-hyang (1994) *A Cross-Linguistic Study of the role of the Jaw in Consonant Articulation.*

McGory, Julie (1997) *The Acquisition of Intonation Patterns in English by Native Speakers of Korean and Mandarin.*

Migge, Bettina Margarete (1998) *Substrate Influence in the Formation of the Surinamese Plantation Creole: A Consideration of Sociohistorical Data and Linguistic Data from Ndyuka and Gbe.*

Parkinson, Frederick B. (1996) *The Representation of Vowel Height in Phonology.*

Peng, Shu-hui (1996) *Phonetic Implementation and Perception of Place Coarticulation and Tone Sandhi.*

Roberts-Kohno, Ruth (2000). *Kikamba Phonology and Morphology*

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

Yoo, Eun Jung (1997) *Quantifiers and Wh-Interrogatives in the Syntax-Semantics Interface.*

Yoon, Jae-Hak (1996) *Tezmporal Adverbials and Aktionsarten in Korean.*

Ohio State University Working Papers in Linguistics No. 54

Varia

Table of Contents

TAGGING PROSODY AND DISCOURSE STRUCTURE IN ELICITED SPONTANEOUS SPEECH

Mary E. Beckman and Jennifer J. Venditti

Abstract

This paper motivates and describes the annotation and analysis of prosody and discourse structure for several large spoken language corpora. The annotation schema are of two types: tags for prosody and intonation, and tags for several aspects of discourse structure. The choice of the particular tagging schema in each domain is based in large part on the insights they provide in corpus-based studies of the relationship between discourse structure and the accenting of referring expressions in American English. We first describe these results and show that the same models account for the accenting of pronouns in an extended passage from one of the Speech Warehouse hotel-booking dialogues. We then turn to corpora described in Venditti [Ven00], which adapts the same models to Tokyo Japanese. Japanese is interesting to compare to English, because accent is lexically specified and so cannot mark discourse focus in the same way. Analyses of these corpora show that local pitch range expansion serves the analogous focusing function in Japanese. The paper concludes with a section describing several outstanding questions in the annotation of Japanese intonation which corpus studies can help to resolve.

1 Introduction

The development of a large spontaneous speech Japanese language corpus under the sponsorship of the Science and Technology Agency is a signal event in the illustrious

history of speech technology in this country. Japanese laboratories have been at the forefront in the development of key parts of current automatic speech recognition (ASR) and text-to-speech (TTS) technology — e.g., the use of variable-length units in concatenative speech synthesis [Sagi88]. Because of such contributions in many laboratories both in Japan and elsewhere, speech technology today is at a stage where two more complex and difficult challenges can begin to be addressed seriously. Large vocabulary ASR systems have good word recognition rates even for continuous speech, and our emphasis now can turn to integrating ASR fully with natural language parsing (NLP) technology in order to try to build complete spoken language understanding systems. Also, the basic algorithms for TTS are now good enough that we can begin to integrate them with NLP technology to design complete spoken language generation systems, to try to generate comprehensible dialogues and not just strings of individually intelligible sentences.

These twin challenges of spoken language understanding and spoken language generation require a larger fund of knowledge about spoken language than we now have. This knowledge should build on the speech science and linguistics of the 20th century, but it must go considerably beyond them. A better understanding of prosody and a better understanding of discourse organization will be key elements of this knowledge. Each of these elements requires that we look closely at spoken language in its normal environment: ordinary communicative interactions of the sort that humans engage in effortlessly every day of their lives. In other words, there is an urgent need for large corpora of spontaneous speech elicited in meaningful tasks such as asking for directions. Moreover, these corpora must be processed in such a way that we can build on our current understanding of prosody and discourse organization. The corpora must be tagged for prosodic categories and discourse elements so that we can use them to train and test better models, capable of mimicking the ways in which human speakers and listeners structure spoken language for easy real-time comprehension.

Of course, processing a large spontaneous speech corpus is difficult and expensive. Unlike segment labels or part-of-speech tags, prosodic elements and discourse structures have not been a central focus of the Linguistic Data Consortium in the United States. (In this respect, the Japanese effort is ahead of the American one.) Although there has been at least one research project aimed on ways to speed up the tagging process [SHBMc], the algorithm and the data on which the algorithm was trained are proprietary. Also, spontaneous speech is not a single type of thing (see [Beck97]), and we have no guarantee that tags and tagging algorithms developed for one type of corpus will generalize to fully cover the elements of interest in a different speech style. To put it another way, tagging of prosody and discourse organization is in its infancy, just as segment labelling was in the 1970s, when the TIMIT database was first being created. Therefore, it is still a time-consuming and expensive process. We will need much more manually annotated speech than we have now before we can have automatic tools comparable to Wightman & Talkin's [WT94] aligner program. In order to take best advantage of our current knowledge, we need to design our corpora carefully. We need to start with a good set of initial hypotheses about the kinds of things that we want to observe, and the kinds of relationships that might exist among the segment string, the prosodic organization, the syntax, and the discourse elements. And we

need to experiment carefully with different corpus elicitation protocols.

This paper is a preliminary progress report on the types of elicitation protocols that we have devised, the tags that we are using to annotate the elicited corpora, and the hypotheses that we have been testing with these corpora concerning the relationship between prosody and discourse organization. In the first two sections of the paper, we will argue in more detail for the need to elicit and tag spontaneous speech, using examples primarily from American English, a language that is prosodically and syntactically quite different from Japanese. In this part, we will also describe a general framework for thinking about discourse organization which has proved useful in understanding the relationship between prosody and discourse structure in English. Then, in the next two sections of the paper, we will turn our attention more fully to Japanese. Here we will describe the tagging system that we have developed for standard (Tokyo) Japanese [Ven95] and describe some more recent research that suggests further improvements to this system. Also, we will discuss the kinds of prosodic and syntactic cues that are used to cue discourse organization in Japanese, at least for the corpora that we have looked at so far. Finally, we will list a few of the unanswered questions that could fruitfully be the topic of concerted investigation using corpora that are being developed now, including the corpus sponsored by the Science and Technology Agency, which is the core of this symposium.

2 Why tag prosody?

Ten years ago, it was still possible to disagree about how important prosody is for speech recognition. A speech scientist arguing for the importance of recognizing prosody could point to strings of phonemes or words such as (1)-(4):

(1) /bIlo/
(2) /kaneokuretanomu/
(3) The old men and women stayed at home.
(4) Yu'u-kun to Mine'yori-kun no oni'isan ni aima'sita.

Without any indication of the prosody, we do not know whether to interpret the string of phonemes in (1) as the preposition *below* or the content word *billow*. The string in (2), similarly, is ambiguous between *kane-o kure; tanomu.* 'Send me money, I beg you.' and *kane-o kureta. nomu.* 'I've received the money, and am drinking.' The sentence in (3) is one of Lehiste's [Leh73] classic examples of a syntactic ambiguity which can be differentiated by the intonational phrasing, and the sentence in (4) from [Eda] is a comparable example from Japanese of a syntactic ambiguity that can be disambiguated by the intonational phrasing (see Figure 1).

A scientist on the other side of the debate could always counter by suggesting that such totally ambiguous strings only rarely occur outside of the laboratory, and in ordinary conversation, the (non-prosodic) context typically provides redundant cues to the intended reading. A further argument for this view is the fact that some of the highest levels of

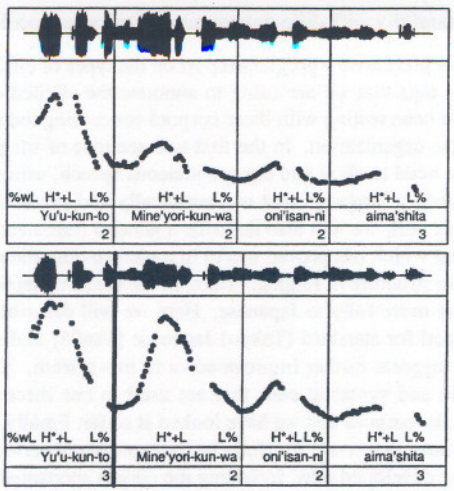


Figure 1: Fundamental frequency (F0) contours and J_ToBI transcriptions of the two readings of the sentence in (4). In the upper panel, the four content words are are all grouped together into a single intonational phrase, and the preferred interpretation is left-branching: 'I met Yuu and Mineyori's older brother.' In the lower panel, there is an intonational phrase boundary between the two proper names (marked with a thick line), and the preferred interpretation is right-branching: 'Yuu and I met Mineyori's older brother.' [Utterances kindly provided by Sanae Eda.]

word-recognition accuracy have been reported for systems that simply plugged the best word models from an ASR system into syntactic models based on text corpora [LR89].

In speech synthesis, by contrast, there has been less room for disagreement. Research on word-level accuracy with non-native speakers [Mack87] and on ease of comprehension in native speakers (e.g., [Sil93]) demonstrated that high word-level intelligibility with native speakers is not a good measure for evaluating TTS systems and that poor prosody makes even the most intelligible synthetic speech difficult to process. More than ten years ago, Klatt [Klatt87] described poor prosody as the single largest contributing factor in the poor quality of even the most highly intelligible synthetic speech of his day, and TTS researchers today still agree with his assessment (see [SOH99]). Moreover, as we move beyond ASR and TTS to spoken language understanding, and generation, the need for good models of prosody becomes increasingly clear.

Figure 2 illustrates this point. It shows transcripts of two extracts from a dialogue elicited using a hotel and airline booking paradigm. Speaker S (Steve) is acting as the travel agent, and is sitting in front of a computer with an online reservation system. Speaker T (Tom) is simulating a client who is talking to S over the telephone. This elicitation

56 S: Uh okay, I uh sorry to say I I don't believe the Best
Western is handicapped accessible. At least(12) the
57 T: Uh huh(12) Okay.
58 Well I have one more choice for you.
59 S: Uh huh?
60 T: That would be the McClure — M C C L U R E, I think.
61 S: Okay, just one(13) minute here while I(14)
62 T: It might(13) Okay(14)
64 S: You say McClure? M C?
66 T: It — and then it's either McClure or McLure.
I'm not sure if there's a 'c' after the first 'c'.
So we might(15) have to try it two ways.
67 S: Okay(15)
68 Well, we'll try it here with M C C L U R E,
would that be?(17)
69 T: Right(17)
70 S: Okay.
71 Well, let's we'll we'll try that and see what a
72 Uh yeah now we don't f- have any listings for
that particular spelling uh(18)
73 T: Okay(18)
74 S: Shall we try the (19) M C L (20) U R E?
75 T: uh(19) Uh huh(20)
76 Uh huh
77 S: Okay, let's try that.
78 Okay, yes. McLure(21) House, Hotel and
Conference Center. Great.
79 T: Good(21)

[S sees that the McLure does not accept online reservations and gives T the toll-free number for the hotel. He then goes on to look up other hotels in the area.]

115 S: There's the Holiday Inn Express is the uh one other
option that we have here.
116 T: Hmmm. I didn't know about that one.
117 S: Uh huh. Yeah this is on I-seventy and Dallas Pike.
118 T: Ah.
119 S: Um, so maybe it's new.
120 T: Well, I think that one's been about five different
chains over the last ten years(24). That's what it is
today. Let's see tomorrow.
121 S: Aha okay(24)
122 S: Now, let's see um. Okay
123 Uh we can reserve rooms here
124 Uh(25) let me check on uh the the types of rooms
that are available.
125 T: Uh huh(25)

Figure 2: Two extracts from the transcript of a hotel booking dialogue. Underlined text indicates overlap with the other participant's turn, and overlapped portions are co-indexed.

paradigm was designed by Julia McGory and Stefanie Jannedy, and we are using it extensively in our current research, because hotel and airline reservations are one domain where spoken language technology could allow ordinary people to access specialized computer databases in a convenient way without having to pay for internet access in their homes. Ideally, the querying system should be able to process the client's intents and respond appropriately, with the same conversational skills that a human travel agent brings to the task. In order to sample these skills, we have elicited dialogues between S and several clients, with diverse travel needs and expertise — i.e., different amounts of local knowledge relative to the agent's. In this particular dialogue, T is returning to his home town for a funeral, needs a room with wheelchair access, and is suggesting various hotels for S to look up.

The extracts in Figure 2 give several examples of the ways in which prosody aids the negotiation of information flow between the two participants in the dialogue. A particularly striking case is utterance 117, where S is giving T information about the Holiday Inn Express, first mentioned in utterance 115. This utterance is syntactically a declarative sentence, and the context makes it clear that T is interpreting it as an assertion of information. Yet the boundary pitch movement at the end is very similar to the rise that is typically associated with a yes-no question (see Figure 3). It is possible to use intonation to mark a syntactic declarative as a yes-no question in English, so this case is worth examining in more detail. The canonical yes-no question intonation in American English is L* H- H% — that is, a large rise from a low pitch target on the last accented syllable (L*) through a high pitch target phrase tone (H-) and on up to an even higher pitched target at the very end of the phrase (the H% boundary tone). Listening to utterance 117, we can hear very clearly that the rise at the end of this sentence is not the 'low rise' of the yes-no question, but something more like the 'high-rise' pattern that Pierrehumbert & Hirschberg [PH90] discuss in arguing that boundary pitch movements should be decomposed into a part that belongs to the boundary per se, and another part that belongs to the last accented syllable. That is, the first part of the rise here can be attributed to the transition from a low target on the *Dallas* to a high pitch accent (H*) on the word with main stress *Pike*. This accent is typically associated with assertions. Thus, S is making an assertion here (as the accent type makes clear), but he is also doing more. The further rise to the H- H% boundary sequence is expressing something like 'Does that sound familiar? Can you identify the hotel with that added information, and will that location serve your needs?' And T's response makes it clear that this is indeed how he interprets S's statement. If the intonation pattern here were not tagged correctly, we would not be able to distinguish the low-rise from the high-rise tune correctly in the way that we should to train a spoken language system to generate the travel agent's turns in exchanges such as this.

Another striking example of why we need to tag prosodic elements in these utterances is the accent pattern in utterances 71 and 77, two places where S says *Let's try that.* The syntax is the same, and in each case *that* is a pronoun referring back to information introduced earlier — i.e., one or the other of two possible spellings of the name *McClure*. But the two utterances differ prosodically (see Figure 4). In utterance 71, S places a pitch accent on the verb *try*, whereas in utterance 77, he accents *that* instead, using the rising (L+H*) pitch accent whose discourse function has been studied by Ladd [Ladd80], Ward

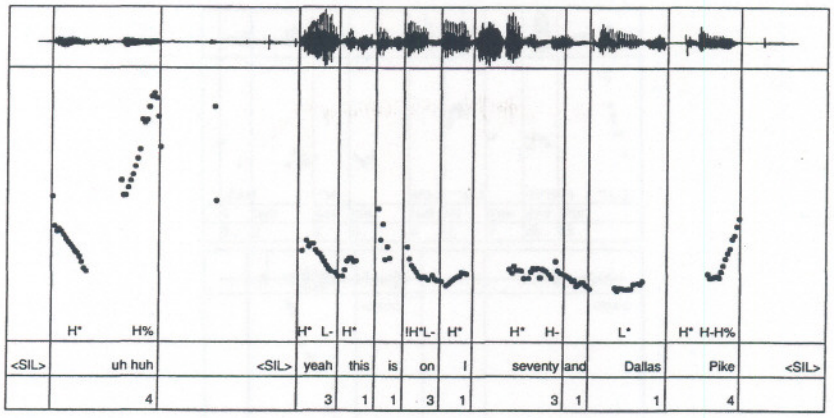


Figure 3: F0 contour and AmerEng_ToBI transcription for utterance 117 from the hotel booking dialog in Figure 2.

& Hirschberg [WH85], and Cahn [Cahn95], among others. A good concept-to-speech system should be able to predict when a pronoun such as *that* will be accented, and also to generate an appropriate pitch accent type for the context. In order to build a good predictive generative model, we need large domain-appropriate spontaneous speech corpora, with utterances tagged for accent pattern and type. (We also need to annotate the corpora for the discourse elements and structures that might help us understand precisely why the accent on *that* is appropriate in one case but not the other, but that is a separate issue, to which we return in the next section.)

As these examples show, boundary pitch movements (such as the rise to a H% intonation phrase boundary tone at the end of *Dallas Pike* in Figure 3) and pitch accents (such as the rising L+H* tone on the pronoun *that* in the lower panel of Figure 4) are prosodic elements that are important to identify accurately in American English spoken language corpora. The tags that we show in Figures 3 and 4 are the American English ToBI (AmerEng_ToBI) labels for intonational events. The AmerEng_ToBI system is based on a large body of work on the prosodic system of English (e.g., [Pierre80, PH90, POSHF91]), and has been demonstrated to have a high degree of intertranscriber consistency (e.g., [PBH94, MHS99]). Currently, the only way to extract these events accurately is to train human labelers to tag them manually. Figure 5 (from [McG99]) illustrates one of the reasons why this is the case.

The upper panel in Figure 5 shows two more rising boundary pitch movements like the one at the end of utterance 117 in Figure 3, but in this utterance, the first rise is in the middle of the utterance, where it is in contrast with the rising pitch accent in the lower panel in Figure 5. The contrast here illustrates another important point about English prosodic structure. The alignment of pitch events relative to the associated text is just

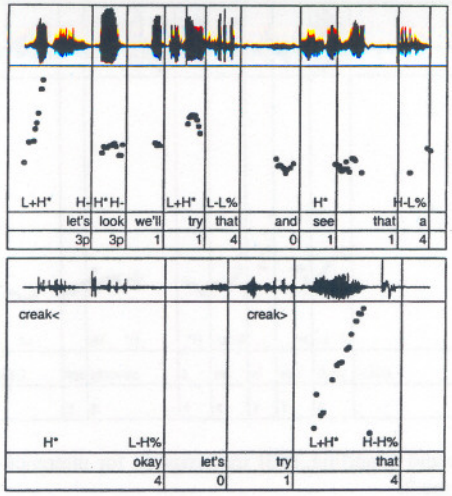


Figure 4: F0 contours and AmerEng_ToBI transcriptions for utterances 71 and 77 from the hotel booking dialog in Figure 2.

as important as the gross pitch shape. The rise fall rise pattern is nearly identical in the two utterances in Figure 5. To the native speaker's ear, however, the difference is quite striking and obvious. The rise in the upper panel marks an intonational phrase boundary, whereas the one in the lower panel marks an accented syllable. Smoothing the F0 contour in an attempt to 'undo' microprosodic effects (as in [Tay93]) will only obscure the subtle intonation differences that do exist in this case. This makes it impossible to extract the relevant prosodic elements from a spoken language corpus on the basis of the fundamental frequency contour alone. Ostendorf & Ross [OR97] attempted to recognize the tune using other cues to phrasing and accentuation as well as the alignment of the F0 contour with the words. Their system had modest success on a read speech corpus in a news-caster's reading style. With enough hand-labeled data in several speech styles, we should be able to generalize such an algorithm to spontaneous speech in other domains where it can be applied fruitfully in a complete spoken language understanding and generation system.

3 How should we tag discourse structure?

Once we have prosodic tags for a spoken language database, such as the dialogue illustrated in Figures 2–4, we can begin to think about predicting the tags from other aspects of the corpus. As Figure 1 suggests, prosodic structure is constrained by the syntactic structure. The relationship was noticed very early in the history of modern linguistics, and there is now a large body of literature relating the two. (See [Selk84] for just one relatively recent

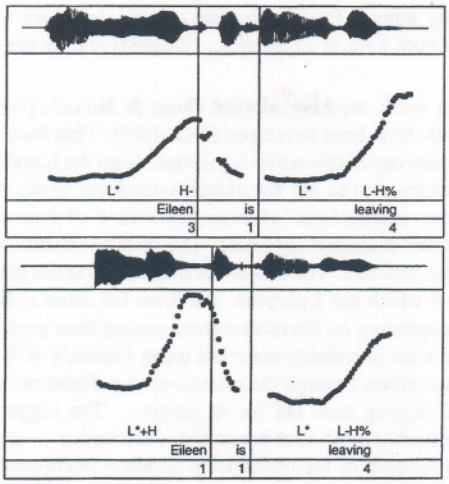


Figure 5: F0 contours and AmerEng_ToBI transcriptions for utterances illustrating two functionally distinct rise-fall-rise patterns. In the upper panel, the rise is an interpolation from a L* pitch accent on *Eileen* to a H- phrase tone at the end of the first of two (intermediate-level) intonational phrases. In the lower panel, the rise is a L*+H pitch accent on *Eileen*, and there is only one intonational phrase. [Utterances kindly provided by Julia McGory].

monograph.) As Figures 3 & 4 demonstrate, however, syntax is far from the only structure that constrains prosody. In order to be able to predict the different boundary shapes in Figure 3 and the different accent placements in Figure 4, we need to look beyond the syntax of individual utterances. We need to have an understanding of the larger discourse context and the ways in which that context is structured. In other words, we need a general framework for describing the discourse structure, and an associated standard system for tagging the elements and features of this particular discourse.

In order to constitute a standard, a tagging system must meet several criteria. It should be built on a body of established knowledge that is large enough to yield some consensus facts (if not a consensus theory to explain the facts). The tags should provide enough coverage of established phenomena that it can be adopted by a reasonably large proportion of the community of potential users. That is, it should fill the intersection of needs across the community. The tags must be specified precisely enough that they can be applied consistently, and training materials should be supplied so that new users can learn the system, and use it to tag a corpus in the same way that a more experienced user does. The last criterion can be established in intertranscriber consistency tests, using standard statistical tests of agreement such as Cohen's kappa (see [Fle71]). It is not as easy to

establish that a tagging schema fits the first two criteria, but there has been attempts to establish a consensus both here in Japan (e.g., [dtag98]) and in the United States (e.g., [acl99]).

In much of our work, we have adopted Grosz & Sidner's [GS86] framework, for which training materials have been developed [NGAH95]. This framework identifies two other aspects of discourse organization that are distinct from the linguistic structure of sentence fragments, sentences, and so on: the global 'intentional structure' of discourse segments and their purposes, and the local 'attentional structure' of dynamically shifting focus states within and between discourse segments. The intentional structure is an unfolding, but ultimately static tree structure. The utterances in a discourse are grouped into discourse segments (DS), each of which has a purpose, and these DS stand in hierarchical relationships to one another, depending on the relationships among their purposes. Nakatani et al. [NGAH95] developed a set of training materials using Flammia & Zue's [FZ95] tagging tool, which guides transcribers through the utterances of a discourse, grouping utterances together into DS, and tagging each DS for its purpose. The tagging scheme has been shown to produce reasonably good inter-transcriber consistency — good enough to allow for a meaningful investigation of the relationship between intentional structure and such intonational properties as phrasal pitch range (e.g., [GH92]).

In our own work ([VS96, Ven00]), we have applied this framework for understanding the relationship between intentional structure and prosody to Japanese, and have found good agreement with the attested results for English, once the differences between the two prosodic systems have been taken into account (see Section 5). This is not surprising, given the general consensus that exists about intentional structure and its relationship to such properties as phrasal pitch range. Indeed, discourse segmentation and the intentional hierarchy has been studied for centuries in the guise of 'rhetoric' and tagging schema for this aspect of discourse organization can build on the everyday skill that a schoolchild exercises when producing a hierarchical 'outline' for an essay or report in elementary school.

By contrast, there has been less clear agreement about how to tag attentional structure. This aspect of discourse organization is related to the theme/rheme division posited by the Prague School linguists, Halliday [Hal67], and others. In much of our work, we have adopted the framework of Centering Theory [GJW95] as our model of attentional structure. In this framework, an utterance has a 'Center' — the focal discourse entity that the utterance is most centrally about. When it is not the first utterance in the discourse, the Center is 'backward-looking' — i.e. it can be identified with one or another candidate entity in a list of 'forward-looking Centers' in the preceding utterance. No standard tagging tool has been developed for Centering Theory. Hence, there are no intertranscriber consistency tests for Centers and Center relationships comparable to those for intentional structure. However, there is consensus among researchers in this framework on criteria for identifying and ranking the forward-looking Centers, and for identifying the backward-looking Center, based primarily on language-specific syntactic criteria (e.g., [WIC94], for Japanese). This has enabled individual researchers to tag some spontaneous speech corpora (e.g., [Naka97, Pass98]), and research using this approach has suggested a way to predict when a pronoun will be accented in English.

The literature on accentuation and its relationship to information status in English predicts that a pronoun typically should be unaccented. That is, a pronoun refers back to an entity which is currently salient in the discourse (i.e., the Center). Therefore, it should not be accented, because it represents 'old' information. Nakatani [Naka97] examined the discourse functions of pitch accent on pronouns in a spontaneous narrative elicited using a standard sociolinguistic interviewing protocol. She concluded that pronouns are generally unaccented when they continue the current Center, while they are accented when they serve to shift the Center of attention to another entity in the discourse.

This generalization is in keeping with the accent patterns in Figure 4. When the pronoun *that* occurs unaccented in utterance 71, it is referring to the spelling with two 'C's, which continues the Center introduced in utterance 68. (Note that the *that* in the last clause of that utterance also is unaccented.) When that occurs accented in utterance 77, by contrast, the Center is shifting to the alternate spelling with only one 'C' (cf. utterance 74). On the other hand, this result obviously cannot generalize to Japanese, because Japanese does not use pronouns in the way that English does. When there is not simple ellipsis (i.e. a 'zero pronoun'), the more standard way to refer to the Center is with a topicalized noun phrase marked with the postposition *wa* (see [WIC94]). Therefore, the relationship between prosodic structure and attentional structure will necessarily be different. Before describing our work on prosodic cues to attentional structure in Japanese, however, we must amplify on another reason why the result does not generalize — the fact that the prosodic function of pitch accent in Japanese is quite different from that of accent in English.

4 The J_ToBI system

Although Japanese is prosodically quite different from English, it is possible to adopt the same general framework for tagging critical prosodic elements. In our work, we have adopted the J_ToBI labelling conventions [Ven95]. The J_ToBI conventions are a method of prosodic transcription for Tokyo Japanese which is consistent with the five general principles adopted by developers of ToBI conventions for other languages. The first of these principles is that the labelling conventions must be "as accurate as possible, given the current state of knowledge. Ideally, they will be based on a large and long-established body of research in intonational phonology, dialectology, pragmatics and discourse analysis for the language variety, but at the very least, they are based on a rigorous analysis of the intonational phonology." (See http://ling.ohio-state.edu/ tobi for these principles, and a list of other languages for which ToBI framework systems have been developed.) The J_ToBI tags are based on a venerable and large body of research on Japanese pitch accent and intonation patterns (e.g., [Hat60, Hat61, Kawa61, Kawa95, Hara77, McC68, PB88, Kubo93, VMvS98, Mae98]).

Among the established facts about Japanese that are reflected in the J_ToBI labels is the lexical contrast between accented and unaccented words. Japanese has pitch accents, much like the pitch accents of English, German, and Greek. For example, in the utterance shown in Figure 6, the words *sa'Nkaku* 'triangular' and *ya'ne* 'roof' are accented, whereas

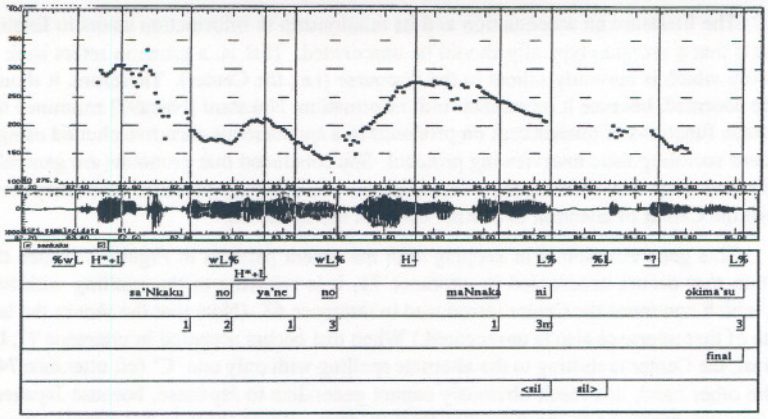


Figure 6: F0 contour and J_ToBI transcription for the utterance fragment *sa'Nkaku no ya'ne no maNnaka ni okima'su.* 'I will place (it) directly in the center of the triangular roof.' [From the J_ToBI Guidelines.]

maNnaka 'center' is unaccented. This difference is reflected in the presence versus absence of the H*+L label marking the accent kernel in the tone tier — the topmost labelling window in the figure. As in the ToBI labelling conventions for English, German, and Greek, the '+' indicates a marker for a pitch accent with two tone targets (the Japanese pitch accent is a fall from a high pitch target to a low one) and the '*' indicates which of the two pitch targets is associated to the accented syllable in the text. Adopting these conventions allows us to capture the essential similarity between pitch accents in all of these languages, a similarity that was noted long ago by Hattori [Hat61], McCawley [McC68], and many other researchers. That is, a pitch accent is a tone pattern that is aligned with a designated (accented) syllable within a word.

At the same time that the ToBI framework captures this cross-language similarity, it also allows us to acknowledge any crucial prosodic differences. Two differences are relevant. First, in Japanese, a pitch accent necessarily causes a 'downstep' — a steplike reduction of the pitch range within the intonational phrase. In the utterance fragment in Figure 6, for example, the first word *sa'Nkaku* is accented. This triggers downstep, so that the accent peak on the second word *ya'ne* is much lower. In the last part of Figure 7, by contrast, the word *heikoo-ni* 'level' is unaccented, and so does not trigger downstep. In this utterance, the accent peak on the following phrase *narabu yo'o ni* 'so as to line up' is nearly at the same level as the highest point in the *heikoo ni*. In English, downstep involves a choice of accent type, and the AmerEng_ToBI labels mark it explicitly, using the '!' diacritic. (See the word *on* in Figure 3.) In the J_ToBI conventions, we do not mark downstep, because it is predictable from the lexical accent.[1]

[1]This is in keeping with the second principle of building ToBI framework systems: "The conventions are

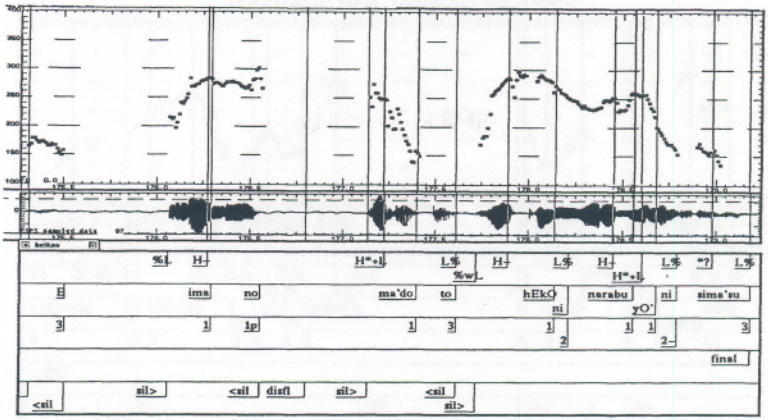


Figure 7: F0 contour and J_ToBI transcription for the utterance *ima no ma'do to heikoo ni narabu yo'o ni sima'su.* 'I will make it so that they line up level with the livingroom window.' [From the J_ToBI Guidelines.]

The second relevant difference between Japanese and English is that pitch accents in Japanese are not associated with 'stressed' syllables (cf. the discussion of accent placement in the utterances in Figure 4 above). There is nothing in a label such as H*+L that necessarily implies that the accented syllable is prosodically prominent. This is as it should be, because the contrast between accented and unaccented words in Japanese has nothing to do with the kind of intonational prominence that governs pitch accent placement in English, German, Greek, and other 'stress-accent' languages. Rather, the placement of pitch accents in a Japanese utterance is governed by phonological specifications inherent to the words themselves. The two accented words in the utterance in Figure 6 are inherently accented; this is part of their lexical specification and not due to any perceived intonational prominence. Indeed, in this utterance, the unaccented word *maNnaka* is perceived as being much more prominent intonationally than the accented word *ya'ne* that immediately precedes it.

Another established fact about Japanese that the J_ToBI prosody tagging conventions capture is the distinction between two levels of intonationally marked prosodic grouping. The first level is the accentual phrase. This level of prosodic constituency is marked canonically by a rise in pitch at the beginning. For example, in the utterance fragment in Figure 6, there is an accentual phrase boundary between *sa'Nkaku no* and *ya'ne no*. Similarly, in the utterance in Figure 7, there is an accentual phrase boundary between *heikoo ni* and *narabu yo'o ni*. This level of phrasing is indicated by the break index value of 2

efficient. They do not waste transcriber time by requiring the transcriber to symbolically mark non-distinctive pitch rises and falls that can be extracted from the signal automatically, or anything else that could be extracted from resources such as online pronunciation dictionaries."

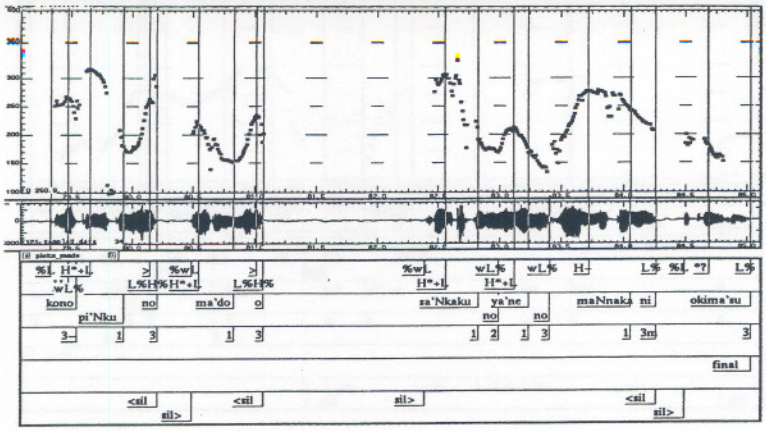


Figure 8: F0 contour and J_ToBI transcription for the utterance *pi'Nku no ma'do o sa'Nkaku no ya'ne no maNnaka ni okima'su.* 'I will place a pink window directly in the center of the triangular roof.' [From the J_ToBI Guidelines.]

on the tier of labels just beneath the romanized transliteration of the words in each figure. Contrast the lack of any pitch rise at the word boundary between *narabu* and *yo'o ni* in Figure 7. These two words are grouped together into the same accentual phrase, as typically happens when a content word such as the verb *narabu* is followed by a function word such as the postpositional adverbial *yo'o ni*. (See [SS83, Kubo93] for studies of this.) Such phrase-internal word junctures are marked by break index *1* on the break index tier.

The other level of intonationally-marked prosodic grouping is the intonational phrase. It is marked in the intonation pattern primarily by a new choice of pitch range — a pitch range 'reset' which undoes any downsteps that have been triggered by accented lexical items in the preceding phrase. In Figure 6, for example, there is an intonational phrase boundary just before *maNnaka*, so that *sa'Nkaku no* and *ya'ne no* are in a separate phrase, and *maNnaka* is not doubly downstepped by the two accents. This phrase boundary is reflected in the break index value of *3* on the break index tier.

Another (optional) pitch event that has been assumed to be a marker for the intonational phrase is the occurrence of 'extra' boundary tones to provide a distinctive 'boundary pitch movement' pattern. This is illustrated in Figure 8, where the first two phrases end with a rising boundary pitch movement, which is accounted for in the tones tier by the rise from the L% that marks the end of the accentual phrase to a following H% at the intonational phrase edge.

Note that the pitch peak on *ma'do* 'window' is lower than the pitch peak on *pi'Nku* 'pink' in the preceding intonational phrase. Looking just at these pitch range relationships in the F0 contour, we might think that the second word is subject to the downstep triggered by the first word — i.e. that *ma'do* does not begin a new intonation phrase after all, despite

the boundary tone. However, native speakers who listen to the audio file tend to agree with the transcription here. The boundary pitch movement gives a clear sense of a disjuncture that is more pronounced than expected for a mere accentual phrase.[2] On the basis of such native speaker judgments, we assume that there is an intonational phrase break here in this utterance. Therefore, we cannot attribute the pitch range relationship to a downstep triggered by the accent on *pi'Nku*. We account for the appearance of downstep instead by saying that while the pitch range has been 'reset', the choice of the new pitch range here is one that subordinates *ma'do* pragmatically to *pi'Nku*.[3]

With this background, we can now explain the perceived prominence on *maNnaka* in Figure 6. The word is prominent because it begins a new intonational phrase, and the choice of the new reset pitch range is a very wide pitch one, so that there is a very pronounced rise in F0 from the L% boundary tone at the end of *ya'ne* to the H- phrase tone that is anchored on the first syllable of *maNnaka*. In other words, while pitch accents in Japanese cannot play an analogous role to English pitch accents in cuing Centering relationships, we can look at pitch range relationships between adjacent phrases as potential cues to what is salient within the discourse segment.

5 Prosody and discourse structure in Japanese

Our current research on Japanese (particularly [Ven00]) focuses on pitch range variation in connected discourse. Our working hypothesis is the following: a great deal of the variation in pitch range observed in connected discourse can be correlated with the same kinds of syntactic and discourse tags that have been used to predict pitch accent distribution in English (e.g., [Hirsch93]).

Figure 9 shows some of our preliminary results, using a database of spontaneous and read monologues. The monologues were elicited using the following protocol (described further in [Ven00]). First a spontaneous monologue is elicited by asking the speaker to narrate a story about two girls meeting in the park. Sequences of hand-drawn pictures were used as prompts. This elicitation method minimizes the memory load on the speaker narrating the story, resulting in a fluent spontaneous discourse containing few hesitations or other disfluencies. Then, after a few spontaneous monologues have been recorded, any later speaker can be recorded also reading a monologue that is the written transcription of one or another of the previously elicited spontaneous monologues. The elicited spontaneous and read speech data are then segmented and tagged using prosodic (J_ToBI) tags, syntactic tags, and discourse structure tags. These tags then are used to analyze the pitch range variation, as in Figure 9.

[2]This illustrates another of the principles of the ToBI framework: "The conventions do not replace a permanent record of the speech signal with a symbolic record. An electronic recording of the transcribed utterance is an essential component of a complete ToBI framework transcription." That is, listeners have access to other cues to the disjuncture, and listening is an essential component of tagging the prosody.

[3]An alternative interpretation is that boundary pitch movements can occur at accentual phrase boundaries internal to the intonational phrase. See [MK00].

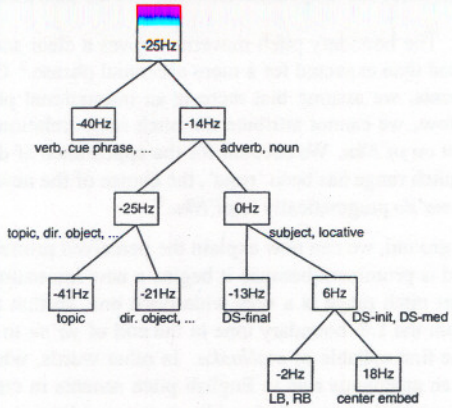


Figure 9: CART tree showing a model of pitch range differences (observed-predicted peak heights) according to tagged features in a read monologue. The tree and features shown here have been truncated to save space.

The figure shows a Classification and Regression (CART) tree which models the pitch range variation in one of the read monologues. Splits in the tree are determined by which combinations of features and feature values will minimize the prediction error after that split (see [Ril89] for a review of this implementation). The hertz value in each square is the average difference between the observed F0 peak value and the peak value that is predicted by our 'default' pitch range model. The default model includes variables such as the amount of reduction at each downstep and typical initial values for the pitch range topline and baseline. These are speaker-specific values, and are extracted for each speaker from a standard set of read sentences. Because the default model accounts for these 'purely phonetic' influences on pitch range, the graphic presentation of the deviation from predicted value in the CART diagram highlights the syntactic and discourse features which are most important for pitch range prediction in this dataset.

There are important deviations from the predicted value, in both directions. Cue phrases (such as *tugi ni* 'next') and verbs are on average produced in a lower range than predicted (the peaks are 40 Hz lower), while adverbs and nouns pattern differently by being produced in a higher range (albeit still lower than predicted by 14 Hz). Among nouns, *wa*-marked topics and objects have a lower range, with topics being realized in a very low range: more than 40 Hz below the predicted value. On the other hand, (*ga*-marked) subjects and locative noun phrases are produced right at the predicted height. Among this subset of noun phrases, NPs that are final to the discourse segment (DS) are lower than DS-initial or DS-medial ones, and NPs located at the left edge of a right-branching center-embedded syntactic construction are realized in a range nearly 20 Hz higher than predicted.

One thing that this analysis shows is that the pitch range of discourse entities in Japanese cannot be accurately predicted from a simple algorithm which uses a single default topline and reference line, along with constant reductions for downstep and unaccented words, even if these values are based on the speaker's own data, as was the case here. There is a large amount of variation in pitch range within sentences and across discourses even after these 'purely phonetic' sources of variation are taken into account. On the other hand, much of this 'extra' variation can be predicted for text-to-speech applications by enriching the text-analysis preprocessing component to tag features such as part of speech. That is, many of the features which cause the pitch range to deviate from the default can be extracted from the text directly.

Another issue that this example brings to light is the marked reduction of pitch range on *wa*-marked topic NPs. Figure 9 shows that topics in this monologue are on average 40 Hz lower than predicted, while other NPs are realized right at the predicted height. Why should topics be realized in such a low range? We hypothesize that this is an effect of both the global and local attentional status of topics in Japanese.

Entities are often introduced into the discourse using a non-topic form, such as NP-*o* or NP-*ga*, and then are referred to again in the same discourse segment with NP-*wa*. In such cases, the *wa*-marked NP is in global attentional focus; that is, it is salient in the current discourse segment. Venditti & Swerts [VS96] report effects of global attentional state on pitch range in Japanese spontaneous housebuilding monologues. In this task, speakers construct the front-view of a house out of geometrically shaped pieces of colored paper. The speakers describe their actions — identifying the piece of paper being used and the part of the house being built — as they perform the task. Venditti & Swerts tagged the data with J_ToBI prosodic labels and a Grosz & Sidner [GS86] style of intentional structure segmentation. They found that discourse entities were realized as 'prominent' (in terms of a relative comparison of pitch ranges) when they were introduced into a discourse for the first time, or when they were re-introduced in a segment after having already appeared in a previous non-adjacent segment. This result is reminiscent of the traditional 'given/new' distinction, here having been replicated with a well-defined notion of discourse structure. This effect of global attentional state on the 'prominence' of discourse entities was also seen in Nakatani's [Naka97] study of English pitch accent distribution. She also found that full NPs are realized as accented when they are introduced or reintroduced into a discourse segment. The difference between the two studies is mainly the definition of prosodic 'prominence': in English prominence is manifested by the placement of pitch accents, and in Japanese by the choice of phrasal pitch range.

In addition to having this global attentional salience, *wa*-marked NPs are often salient in the local context as well. Topics signal what is currently being talked about in the discourse, and as such can often be equated with the discourse Center (e.g., [WIC94]). Where English uses unaccented pronouns to cue the Center, Japanese uses either zero pronouns or *wa*-marked NPs. In the case of zero pronouns, there is of course no acoustic means to mark this local attentional salience, but on NP-*wa* forms, the salience status of the Center is cued by a reduced pitch range. That is, whereas in English, discourse entities that are already currently in local focus are realized by non-prominent (unaccented)

pronominal forms, in Japanese the cue that an expression refers to an entity already in local focus is the choice of a non-prominent (i.e. reduced) pitch range on a *wa*-marked form. Nakatani [Naka97] and Cahn [Cahn95] describe how, in English, a pitch accent on a pronoun can serve to cue a shift in discourse Center to another globally salient entity. Recent results from [Ven00] indicate that expanded pitch range on NP-*wa* forms in Japanese can serve the same function: they cue a shift in discourse Center.

In summary, it is clear that variation in placement of pitch accents in English or choice of pitch range values in Japanese is something that linguistic and computational models of spoken language need to address. The variation is not random, but can be predicted to a large extent by lexical, syntactic, and discourse properties of the speech. It is only with a principled method of tagging prosody, discourse and other linguistic structures, coupled with a large tagged speech corpus, that we will be able to advance our understanding of this systematic variation of prominence markers in spoken discourse.

6 Where do we go from here?

We introduced the work described in the previous four sections by calling this paper a 'preliminary progress report'. We used this term to remind ourselves that research using tagged corpora is an iterative process. For every initial question that is answered, new issues arise. Some of these issues can be investigated with new analyses of the same corpora. Others require us to record new corpora whose design requirements become clear only as we work on already tagged corpora. There are also inevitably questions that arise about the tagging systems themselves. We have already touched on some of these issues and questions in describing the work above. In this section, we close by listing two more of the outstanding questions for Japanese speech corpora.

The first involves the inventory of ways to end an intonational phrase. Currently, the J_ToBI conventions distinguish only three types of boundary tone for the end of the intonational phrase. However, Kawakami [Kawa95] described five types of boundary pitch movements, and more recent work by Venditti and colleagues [VMvS98, Ven99] and Eda [Eda] confirms that there are more types than can be distinguished by J_ToBI tags. The examples in Figure 10 (from [Ven95, Ven99]) illustrate two different rising boundary pitch movements that Eda [Eda] shows to be categorically distinct for native listeners of Tokyo Japanese. In a current collaboration with Kikuo Maekawa, we are working to incorporate the results of this more recent work on boundary pitch movements into the J_ToBI tagging scheme. Corpus studies would be useful for examining the distinctions further. To undertake these studies, however, we need to design elicitation protocols for types of spontaneous speech that might yield instances of the two different types of rises shown in Figure 10, the second of which is not at all typical of read lab-speech styles.

Another question arises from the way that the J_ToBI tagging scheme distinguishes accented and unaccented phrases. Recall that these are distinguished by the presence versus absence of the H*+L marking the accent kernel. This implies that the fall at the accent is prosodically independent of the rise at the beginning of the accentual phrase. In Fujisaki's

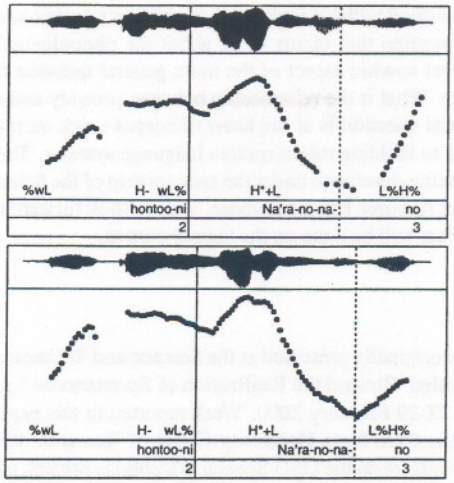


Figure 10: F0 contours and J_ToBI transcriptions of two readings of the sentence *hontoo ni Na'ra no na no*. In the upper panel, the sentence is produced as a yes-no question ('Is it really the one from Nara?') whereas in the lower panel, it is a particularly insistent declarative ('It is really the one from Nara, and that's that!'). The dotted line marks the onset of the final particle *no*.

[FS71, FH84] model, by contrast, the accent fall is a mirror image of the phrase-initial rise, once an automatic and fixed declination of the phrase's pitch range reference line has been factored out. While our default pitch range prediction model (described in the previous section) does not have an automatic fixed declination at the accentual phrase level, it is like Fujisaki's model in linking the size of the accent fall to the size of the rise at the beginning of the accentual phrase. It does this by specifying a (variable) local topline for each accentual phrase, and then fixing the targets for both the H- tone at the beginning of all phrases and the H*+L peak in all accented phrases relative to this same topline. In our corpus work, however, we have seen cases where the H*+L target is clearly higher than the preceding phrasal H- and other cases in which it is clearly lower than the H- target. This variation cannot be predicted by a model in which the relationship is fixed by a constant declination component (as in Fujisaki's model, [FS71, FH84]) or by a fixed relationship to a phrase-level topline (as in our model). A properly designed corpus would allow us to study the relationship between the two high targets, looking at the potential contributions of intervening morpheme boundaries and the syntactic relationships between the morphemes, or the presence of intervening word boundaries and the discourse status of the two words that are grouped together in the accentual phrase.

In other words, the relationship between the rise and fall in an accented accentual

phrase cannot be understood without looking at the phrase's syntax and its role in the discourse structure. A question that seems to be about the phonological model for H tone target turns out to be yet another aspect of the more general question that we asked at the beginning of the paper: What is the relationship between prosody and discourse organization? This more general question is at the heart of corpus work on spoken language corpora, and it is essential to building robust spoken language systems. The large spontaneous speech corpus that is being developed under the sponsorship of the Science and Technology Agency is an important resource for this purpose, and we look forward to seeing the results of the many analyses that will be done on the tagged corpus.

Acknowledgments

This paper was originally presented at the Science and Technology Agency International Symposium entitled "Toward the Realization of Spontaneous Speech Engineering", held in Tokyo, Japan, 28-29 February 2000. Work reported in this paper was supported in part by a grant from the Ohio State University Office of Research, to Mary E. Beckman and co-principal investigators on the OSU Speech Warehouse project, and by an Ohio State University Presidential Fellowship to Jennifer J. Venditti. We are grateful to Julia T. McGory and Pauline Welby for their copious help in preparing the materials from the English hotel booking dialogue and to Julia McGory and Sanae Eda for letting us use examples from their work in Figures 1 and 5.

References

[dtag98] The 3rd workshop of the Discourse Resource Initiative, 1998. Chiba, Japan.

[acl99] Association for Computational Linguistics Workshop: Towards Standards and Tools for Discourse Tagging, 1999. College Park, Maryland.

[Beck97] Beckman, Mary E. 1997. A typology of spontaneous speech. In Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi, editors, *Computing Prosody*, pages 7–26. Springer-Verlag, New York.

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

[Eda] Eda, Sanae. (submitted). Discrimination and identification of syntactically and pragmatically contrasting intonation patterns by native and non-native speakers of Standard Japanese. *Applied Psycholinguistics*.

[FZ95] Flammia, Giovanni and Victor Zue. 1995. Empirical evaluation of human performance and agreement in parsing discourse constituents in spoken dialogue. In *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1965–1968, Madrid, Spain.

[Fle71] Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

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

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

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

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

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

[Hal67] Halliday, M. A. K. 1967. *Intonation and Grammar in British English*. Mouton, The Hague.

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

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

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

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

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

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

[Klatt87] Klatt, Dennis H. 1987. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America (JASA)*, 82:737–793.

[Kubo93] Kubozono, Haruo. 1993. *The Organization of Japanese Prosody*. Kuroshio Publishers.

[Ladd80] Ladd, D. R. 1980. *The Structure of Intonational Meaning: Evidence from English*. Indiana University Press.

[LR89] Lee, Kai-Fu and Raj Reddy. 1989. *Automatic Speech Recognition: The Development of the Sphinx Recognition System*. Kluwer Academic Publishers.

[Leh73] Lehiste, Ilse. 1973. Phonetic disambiguation of syntactic ambiguity. *Glossa*, 7:106–122.

[McC68] McCawley, James D. 1968. *The Phonological Component of a Grammar of Japanese*. Mouton.

[McG99] McGory, Julia T. 1999. Course materials for Linguistics 795T: Practicum in Intonational Analysis and Labeling. Ohio State University.

[MHS99] McGory, Julia T., Rebecca Herman, and Ann Syrdal. 1999. Using tone similarity judgments in tests of intertranscriber reliability. In *Journal of the Acoustical Society of America (JASA)*, volume 106, page 2242.

[Mack87] Mack, Molly. 1987. Perception of natural and vocoded sentences among English monolinguals and German-English bilinguals. In *Journal of the Acoustical Society of America (JASA)*, volume 81.

[Mae98] Maekawa, Kikuo. 1998. Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia.

[MK00] Maekawa, Kikuo and Hanae Koiso. 2000. Design of spontaneous speech corpus for Japanese. In *Proc. of the Science and Technology Agency Priority Program Symposium on Spontaneous Speech: Corpus and Processing Technology*, Tokyo, Japan, pages 70–77.

[Naka97] Nakatani, Christine H. 1997. *The computational processing of intonational prominence: A functional prosody perspective*. PhD thesis, Harvard University.

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

[OR97] Ostendorf, Mari and K. Ross. 1997. A multi-level model for recognition of intonation labels. In Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi, editors, *Computing Prosody*, pages 291–308. Springer-Verlag, New York.

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

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

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

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

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

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

[Ril89] Riley, Michael D. 1989. Some applications of tree–based modelling to speech and language. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 339–352.

[Sagi88] Sagisaka, Yoshinori. 1988. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 449–452.

[SS83] Sagisaka, Yoshinori and H. Sato. 1983. Secondary accent analysis in Japanese stem-affix concatenations. *Transactions of the Committee on Speech Research S83-05*. The Acoustical Society of Japan.

[Selk84] Selkirk, Elisabeth O. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press, Cambridge, MA.

[Sil93] Silverman, Kim. 1993. Assessing the contribution of prosody to speech synthesis in the context of an application. Paper presented at the ESCA Workshop on Prosody, Lund University.

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

[SHBMc] Syrdal, Ann, Julia Hirschberg, Mary Beckman, and Julia T. McGory. (submitted). Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*.

[Tay93] Taylor, Paul A. 1993. Automatic recognition of intonation from F0 contours using the rise/fall/connection model. In *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Berlin.

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

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

[Ven00] Venditti, Jennifer J. 2000. *Discourse Structure and Attentional Salience Effects on Japanese Intonation*. PhD thesis, Ohio State University.

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

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

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

[WH85] Ward, Gregory and Julia Hirschberg. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 61:747–776.

[WT94] Wightman, Colin and David Talkin. 1994. The Aligner: A system for automatic alignment of English text and speech. Document version 1.7, Entropic Research Laboratory.

PRESUPPOSITION RESOLUTION WITH DISCOURSE INFORMATION STRUCTURES[1]

Paul C. Davis[2]

Abstract

An approach to resolving a number of presuppositional phenomena, including definite descriptions and pronominal anaphora, is described within the larger context of an architecture for query-based, task-oriented human/computer dialogue. The model of discourse context employed assumes that discourse structure is organized around a stack of questions under discussion, which plays an important role in narrowing the search space for referents and other presupposed information. The algorithms of individual presuppositional operators for maintaining discourse structures are presented and illustrated in several example dialogues in which human users interact with an agent in order to make hotel reservations. The overall architecture is compared to SDRT (Segmented Discourse Representation Theory), in terms of efficiency and ease of implementation.

1 Introduction

In the many theoretical treatments of discourse, a number of approaches have been used, some including such features as elaborate discourse structures, vast numbers of rhetorical relations, and plan inference engines. Clearly, as demonstrated

[1]This work is an extension of work presented in Kasper, Davis, and Roberts (1999). My thanks to Bob Kasper and Craige Roberts for their work on the project and earlier paper, and, along with Carl Pollard, for many helpful comments on this draft. Thanks also to Harry Bliss and Will Thompson of Motorola.

[2]Paul C. Davis is the recipient of a Motorola University Partnerships in Research Grant.

in the literature, many of these facets of the theories are necessary. However, in human/computer dialogue, when the domain goal is constrained to a single, overarching task to be completed (such as making a hotel reservation), a number of such theoretical prerequisites can be either simplified or factored out, without greatly reducing the coverage of the system. The goal of such a simplified approach is to make a dialogue system more computationally tractable and efficient (and may also make the system more modular and easier to implement).

In this paper, we begin by presenting a computational architecture for human/computer dialogue and demonstrating how it can be employed to solve a number of presupposition resolution problems in discourse. A highly structured discourse model, in conjunction with a treatment of referring expressions as presuppositional, enables us to develop a common strategy for resolving a number of reference resolution problems, such as pronominal anaphora and definite descriptions. This approach extends to a larger group of phenomena which we take to be presuppositional, including domain restriction, ellipsis, and lexically and syntactically triggered presuppositions. All of these constructions are presuppositional in a broad sense, in that their use assumes that certain information can be retrieved from the discourse context. Thus, we are deliberately adopting a broader sense of presupposition than has been conventionally assumed.[3] After the presentation of our approach, we describe its similarity to a more complex and fully-developed theory, SDRT (Segmented Discourse Representation Theory), and attempt to show how our simplified, modular architecture eases the processing task. The architecture consists of a number of modules. We will be focusing on two key components for the major part of the paper, however, the full system will be described in the final section, where it is shown in Figure 7. The two key components, the Question Under Discussion stack (QUD) and the Common Ground (CG), are central for the discussion in this paper. The QUD offers a means to represent the hierarchical nature of the discourse and provides a way of relating utterances to one another–i.e., for keeping track of which utterances are subquestions of earlier utterances, and which are answers–as well as keeping track of the most immediately salient discourse entities. Representing the hierarchical structure is important for certain problems in dialogue (see the example dialogues below) and the QUD can be useful in constraining the search space during the resolution process. The CG is a record of the informational content of the discourse, as well as what would be assumed to be everyday knowledge of the domain. Both the QUD and the CG may be accessed by what we will term presuppositional *operators* (such as the definite description operator) during the resolution process, and both data structures are necessarily dynamically updated as the discourse progresses. We believe this approach is well-suited to certain genres of task-oriented dialogue, in particular for mixed-initiative query systems, i.e., systems where either the human or the com-

[3]The idea that anaphora and presuppositions are closely related is not new (cf. van der Sandt (1992)). However, our treatment of presuppostions is integrated with our discourse model in a different way from van der Sandt, and we apply it to a broader range of phenomena (although van der Sandt (1999) does extend his approach to include domain restriction).

puter may pose questions to the other, toward the end of completing the dialogue task, which itself is generally constrained to one main goal (such as making a hotel reservation, or ordering a pizza).[4]

We will illustrate our approach with four example human-computer dialogues, shown below. SYS indicates the utterances spoken by the computer system. (Dialogue 1) (discussed in detail in §4.1) illustrates a case of pronominal anaphora resolution (*it* in (8)), in which recognizing the hierarchical structure of the discourse is crucial for identifying the antecedent, which was introduced many utterances earlier. The overall topic of the conversation is the question of where the user can find a hotel for June 15th in New York, and this super-question both facilitates and constrains the interpretation of *it* in (8). This example is similar to the well-known examples of long-distance anaphora in task-oriented dialogues described by Grosz (1981). Our approach is consistent with previous research that uses the intentional structure of discourse to determine a set of potential antecedents for pronominal anaphora. The following dialogues will illustrate how a broader range of reference and presuppositional constructions may also be addressed by using the discourse structure to guide the search for relevant information.

(Dialogue 1)	1)	USER:	I'm looking for a hotel for June 15th in New York.
	2)	SYS:	What part of the city would you prefer?
	3)	USER:	Manhattan, near Central Park.
	4)	SYS:	How many nights?
	5)	USER:	Just 1.
	6)	SYS:	Will anyone be traveling with you?
	7)	USER:	No.
	8)	USER:	Oh, I want **it** to have a swimming pool too.

(Dialogue 2) (discussed in §4.2) shows a definite description, *the hotel* in (7), whose referent can only be uniquely determined with respect to the indefinite hotel description (*a hotel close to Madison Square Garden*) in the question under discussion (1):

(Dialogue 2)	1)	USER:	I want to make a reservation at a hotel close to Madison Square Garden.
	2)	SYS:	What dates will the reservation be for?
	3)	USER:	March 3rd and 4th.
	4)	SYS:	Would you like a single room?
	5)	USER:	Yes.
	6)	USER:	Also, I'll need a conference room on the 4th.
	7)	USER:	I'd prefer it if **the hotel** had one.

[4]It remains an open question whether this approach is a useful one for more open-ended systems/tasks.

(Dialogue 3) (see §4.3) involves a contextually determined domain restriction, with a quantificational determiner *every*, illustrating that domain restriction must be handled in a similar way for a broader class of expressions than those which are normally regarded as referring expressions or presupposition triggers.

(Dialogue 3)
1) USER: Does **the Holiday Inn** have any vacancies for
   a) Tuesday, 12/4 - Friday 12/7?
   b) Thursday, 12/6 - Saturday 12/8?
2) SYS: Yes, several.
3) USER: Do they have a breakfast buffet **every morning**?
4) SYS: a) Yes, Monday through Friday.
   b) No. There's a breakfast buffet Monday through Friday, but none on Saturday.

Finally, in (Dialogue 4) (see §4.4) we give a glimpse into our larger research program, where an elliptical question (3) must be resolved with respect to the question under discussion, in addition to establishing the reference of the definite description *the Marriott*, where the context might contain more than one hotel with that name:

(Dialogue 4)
1) USER: Which hotels near the airport have vacancies?
2) SYS: The Holiday Inn and Sheraton have vacancies.
3) USER: How about **the Marriott?**
4) SYS: No, the airport Marriott doesn't have any vacancies.

The remainder of the paper is organized as follows. In section 2 we discuss our assumptions about the structure of discourse and the related background literature. In section 3, we present the individual operators and algorithms which we have developed in a partially completed implementation of a natural language dialogue system where users interact with an automated hotel reservation booking system. In section 4, we discuss the use of the operators and discourse structures to resolve the reference and presupposition problems shown in the above dialogues. In section 5, we describe SDRT, and then compare the two approaches. In the final section, we present an overview of the complete system and our plans for future development.

2 Background: Discourse Structure

Our characterization of the structure of discourse is based on the general theoretical framework of Roberts (1996), where discourse is formally viewed as a game of intentional inquiry.[5] As in Grosz & Sidner (1986), discourse is organized by the

[5]The material in this section is largely unchanged from that in Kasper *et al.* (1999), and was originally written by Craige Roberts.

interlocutors' goals and intentions and by the plans, or strategies, which the interlocuters develop to achieve them. Following Stalnaker (1979), the primary goal of the language game is communal inquiry: Interlocutors attempt to share information about their world, and the repository of that shared information is the interlocutors' *common ground* (CG). The set of acceptable moves in the game are defined by the (conventional and conversational) rules of the game, and are classified on the basis of their relationship to the goals. Ignoring imperatives, there are two main types of moves (see also Carlson 1983): questions and assertions. If the interlocutors accept a question, this commits them to a common discourse goal of finding a satisfactory (asserted) answer: Like the commitment to a goal in Planning Theory, this strong commitment persists until the goal is satisfied or is shown to be unsatisfiable. An accepted question becomes the immediate topic of discussion, the *question under discussion.* An assertion is a move which proposes an addition of information to the CG.

Roberts defines the structure of a discourse at a given point, its *Information Structure*, as a tuple which includes (among other things) the (totally) ordered set of moves in the discourse (M), CG,[6] and the stack of the questions currently under discussion at that point (QUD). The QUD is ordered by order of utterance and is updated in a stack-like fashion,[7] with questions popped when they are answered (or determined to be practically unanswerable). The ordered set of questions under discussion corresponds to the hierarchical intentional structure of the discourse. The QUD in this structure constitutes the set of *discourse goals* of the interlocutors; the discourse goals are only a subset of the set of common goals of the interlocutors, their *domain goals*, and the discourse goals are subordinate to the domain goals. Hence, the requirement that interlocutors stick to the question under discussion is just an instance of the more general commitment to plans; and in turn, in a fully integrated theory we would expect that domain goals and plans would influence interpretation as directly as the discourse goals represented by the questions under discussion.[8]

Any move in a discourse game is interpreted with respect to the Information Structure of the discourse at that point. There are two main aspects to the interpretation of any given move: its *presupposed content* and its *proffered content*, the latter including what is asserted in an assertion and the non-presupposed content of questions and commands. When an utterance presupposes a proposition p, then in order for the utterance to be felicitous in the context, p must be entailed by the CG (Stalnaker 1979). But in addition, any move in a discourse is interpreted by interlocutors under the Gricean meta-presupposition of Relevance, with Relevance

[6]Formally, in Roberts's (1996) framework, the CG is a function from M to sets of propositions, yielding for each move the common ground of the domain of discourse as it existed just before the utterance which the move represents occurred.

[7]However, all elements of the QUD list are accessible during the interpretation of an utterance. Only the top element is writable, but any entry is readable.

[8]Whether these domain goals need to always be computed will be discussed later in the paper (see sections 5 and 6).

formally defined in Roberts' framework, as follows:[9]

(1) A move m is **Relevant** to the question under discussion q iff (i) m is an assertion such that CG$\cup\{m\}$ entails a partial answer to q, or (ii) m is a question whose complete answer contextually entails[10] a partial answer to q.

(1(i)) tells us that the interpretation of an assertion will be constrained so as to yield a partial answer (possibly via contextual entailment) to the question under discussion. (1(ii)) tells us that the QUD in a felicitous Information Structure is constrained by Relevance so that each question on the QUD must address the (prior) question below it on the stack. Of course, (1) correctly predicts a variety of classical Gricean conversational implicatures, now characterizable as contextual entailments. But Roberts argues that Relevance is also crucial in *presupposition resolution*, broadly construed to include anaphora resolution, the interpretation of ellipsis, and domain restriction (Roberts 1995), as well as lexically and syntactically triggered presuppositions.

We will also assume the general approach to anaphora resolution argued for in Roberts (1999). The CG is augmented with a set of discourse referents familiar to the interlocutors, the *Domain* of the discourse context.[11] All definite NPs, including pronouns and demonstratives as well as definite descriptions using *the*, presuppose both *weak familiarity* and *informational uniqueness*. Weak familiarity (cf. the slightly different notion of familiarity in Heim 1982) is the theoretical realization of anaphoricity, and is licensed by existential entailments of the common ground, not requiring an explicit NP antecedent or even perceptual salience of the intended referent:

(2) **Weak Familiarity**: A discourse referent i is weakly familiar in a context C ($i \in$ Domain(C) and C encodes the information that i has properties $P_i, \ldots, P_k$) iff the Common Ground of C entails the existence of an entity with properties $P_i, \ldots, P_k$.

[9]This definition depends on defining partial and complete answers as is done in Roberts (1996), which is based on Groenendijk & Stokhof (1984)

[10]The notion of *contextual entailment* follows straightforwardly from Groenendijk & Stokhof's (1984) notion of *pragmatic implication*. That a contextually entails b simply means that the union of a with the context (in the present theory, this is the common ground) entails b.

[11]In the implementation, this *Domain* is implicit in the CG, in the sense that for all discourse referents there is an *instance* in the knowledge base, where an *instance* is simply a database object. We use the terms CG and knowledge base interchangibly in the paper, but it is important to realize that the latter is the implementation of former, and further, that there is more than one knowledge base in the system, i.e., one representing the CG and another representing the knowledge that the computer system has about hotels, and the like (e.g., how many rooms are vacant in a given hotel–such knowledge proves crucial when accommodation is necessary). These knowledge base distinctions are made where relevant elsewhere in the paper.

Informational uniqueness only requires that the discourse referent which satisfies the definite's familiarity presupposition be unique among the discourse referents in the local context in satisfying the definite's descriptive content. In other words, a referent need not necessarily be unique in the entire CG, but rather be unique in the current unit of discourse structure. The constraints of weak familiarity and informational uniqueness suffice to characterize the presuppositional content of definite descriptions:

(3) **Presuppositions of Definite Descriptions** (informal): Given a context C, use of a definite description NP_i presupposes that there is a discourse referent weakly familiar in C which is the unique weakly familiar discourse referent which satisfies the (possibly contextually restricted) descriptive content of NP_i.

Unlike Russell's (1905) theory, this does not generally entail semantic uniqueness, although in certain special contexts it will yield the same effect via pragmatic means. Definite descriptions may have their descriptive content contextually enriched in the same way that domain restriction works for operators generally, i.e., via Relevance to the question under discussion. This will be illustrated in our discussion of (Dialogue 4) (in §4.2). Many apparent counter-examples to the presupposition of uniqueness for definite descriptions are explained by appeal to this principled contextual enrichment, as discussed at length in Roberts (1999). Pronouns carry the additional presupposition of maximal salience:

(4) **Presuppositions of Pronouns** (informal): Given a context C, use of a pronoun Pro_i presupposes that there is a discourse referent i in C which is the unique weakly familiar discourse referent that is both maximally salient[12] and satisfies the descriptive content suggested by the person, number, and gender of Pro_i.

This amounts to an additional, conventional restriction on the search space for pronominal antecedents, implemented along the general lines suggested by Grosz & Sidner, and explains the differential distribution of pronouns and definite descriptions. We will discuss how maximal salience is implemented in terms of the QUD stack in §4. These presuppositional constraints result in a straightforward theory of anaphoric reference which explains a broad range of data and can be extended to a treatment of demonstrative NPs as definites, as well.

[12]The notion of the relative salience of discourse referents is discussed at length in Roberts (1998); for the purposes of this paper we assume the existence of an algorithm for ordering referents in terms of their salience. Informally, salience can be thought of as a measure of how closely related the referent is to the current discourse, i.e., how relevant it is, and to what degree it is in the *attention* (in the sense of Grosz & Sidner (1986)) of the participants.

3 The Resolution Process

In Figure 1 we show a simplified version of the main algorithm of the overall dialgoue system, and in Figures 2, 3 (shown in §4.1), 4 (§4.2), 5 (§4.4), and 6 (§4.4), we show simplified, pseudo-coded versions of the algorithms for some of the individual operators. The format of the individual operator terms, *OP(VAR, RESTR, NS)*, shown in Figures 2-6, where OP is the operator, VAR the variable, RESTR the restriction, and NS the nuclear scope, is in the familiar generalized-quantifier style, but may generate some confusion. First, our use of *restriction* and *nuclear scope* is not the same as that sometimes used when speaking of universal and existential quantifiers. When we refer to nuclear scope, we are not refering to the delimiting of the scope of a variable (i.e., we are not refering to the range in which a variable may be legally referred to–the range in which it is bound). Rather, the nuclear scope refers to part of the content of the utterance (i.e., the proferred content of that part of the utterance that the operator term represents). The restriction also refers to part of the content of the utterance (the pressuposed part). Informally, the restriction can be thought of as restricting what it is that is being talked about, while the nuclear scope is what is said about that restricted entity. So for the sentence *The man kissed Mary*, we might have a term such as: $def[x, man(x), kissed(x, M)]$, where the restriction is that x is a Man, and the nuclear scope is that x kissed Mary (for this discussion we ignore the familiarity and uniqueness presuppositions). We carry this format to our treatment of all operators. Thus, for pronouns (see Figure 3), which might traditionally be thought of as variables themselves (and therefore, under this traditional view, it would make no sense for them to have restrictions or nuclear scopes), the restriction again refers to what kind of entity it is, and the nuclear scope refers to what is said about the entity (and does not in any way delimit the scope of the variable), so similar to the definite description case, for *He kissed Mary*, our representation might be $pro[x, male(x) \wedge singular(x) \wedge third\_person(x), kissed(x, M)]$ (again, ignoring the familiarity and uniqueness presuppositions). Thus, we have a format for all of the operators, presuppositional and nonpresuppositional, which gives us a uniform way of delineating what the presupposed content is and what the proferred content is, which is very important for the resolution process.

Together, the algorithms for these operators drive the presupposition resolution process. Of central importance in this process is the maintenance of the QUD stack. Each entry on the stack is represented by a Question Data Log (QDL), an ordered triple which contains the utterance's[13] logical form (ULF), its Contextually Understood Logical Form (CULF), and a set of current discourse referents (CDRS). QDLs represent information about units of discourse structure which roughly correspond to the discourse segments developed by Grosz and Sidner.

Process_utterance (Figure 1) is the top-level function invoked for each discourse utterance.[14] The utterance is parsed to yield a logical form representing its context-

[13] An utterance is a full sentence or a fragment (e.g., "Yes."), and is not, in general, an entire turn.

[14] Of course, many of these steps may be eliminated when the system is the speaker. For example,

independent meaning (ULF). This ULF is further processed by *determine_CULF*, the goal of which is to produce a refined logical form (CULF) and a set of discourse referents (CDRS) by resolving presuppositions with respect to the current context. Presuppositions are represented in the logical form by certain operators, including *def*, *pronoun*, λ (for wh-questions), and *WH_Ellipsis*. The terms introduced by these operators, as well as other generalized quantifier terms, are processed by evaluation algorithms[15] for each operator (we call each of the evaluation algorithms *resolve_term*), each of which (again, see Figures 2-6) encodes individual pressuppostional requirements. The operators evaluate themselves relative to the discourse context. We believe this object-oriented methodology is suitable for implementing and testing different theoretical approaches, yet provides a common framework for development.

The set of presuppositional operators shown covers the examples that we will discuss, but is not intended to be exhaustive (an algorithm for non-presuppositional operators, such as indefinites, is shown in Figure 2). After *resolve_term* has processed a presuppositional term, the variable that it binds will appear on the CDRS list, and will either be identified with a set of referents from the common ground or be unanchored (indicated by '?'). This set of referents is a set of possible alternative referents, and may be required to be a singleton (e.g., in the case of the uniqueness presuppositions of pronouns and definite descriptions, see Figures 3 and 4), or may have more than one element, as in the case of a wh-question (where, for example, there can be many possible hotels given a question such as *Which hotels ... ?*). Once the CULF and CDRS are determined, the discourse structures, including the CG and QUD, are updated, depending on the type of conversational move (i.e., assertion or question). After the dialogue model has been updated, the CULF[16] is sent to the back-end application (e.g., to query or update its database), and the system may generate utterances as needed.

---

the system may pass a logical form directly to the dialogue system, rather than requiring parsing. Alternatively, the ouput of the generator may be fed directly back to the dialogue system for parsing. In this sense a system utterance may be treated no differently from a user utterance, should this behavior be desired. While such a strategy might seem somewhat perverse, it might be used in generation, for example, where different alternative system utterances could be generated and then reparsed, in an attempt to see which are easier to deal with (e.g., which lead to less potential ambiguities), before actually generating the sentences to the user.

[15]We generalize from the more familiar notion of a function to an operator which will have specialized implementation for arguments of different types, which is the general strategy in object-oriented program design. Thus, each individual operator has its own evaluation (or perhaps, more aptly, *resolution*) algorithm, called *resolve_term*. The appropriate version of *resolve_term* is indicated by the type of the operator.

[16]Again, this is a simplification. The application actually receives an operator free CULF, where instances from the CDRS have been substituted for variables. The reason for this is obvious, the application relies on the dialogue system to take care of the resolution process, and has no use for the presuppositional operators.

```
process_utterance (U)
%%% 1. Determine contextually interpreted meaning.
ULF = parse(U)[17]
(CULF, CDRS) = determine_CULF(ULF)

%%% 2. Update discourse structures.
If presuppositional operators remain,
   indicate non-acceptance of move (resulting in a prompt for clarification)
If U is an assertion:
   assert CULF to CG,[18]
   update QDL of QUD[top] (i.e., merge CDRS into CDRS of QDL)
If U is a question:
   push new QDL <ULF, CULF, CDRS> onto QUD

%%% 3. Signal back-end application.
Perform SYSTEM action (e.g., query or update database)
Perform SYSTEM dialogue move if necessary (e.g., generate a response)

------------------------------------------------------------------------

determine_CULF (U)
  if atomic_formula(U)  % contains no operators
     return (U, {})
  else (U must contain an operator)
     return resolve_term(U)
```

Figure 1: Dialogue system driver algorithm & determine_CULF

The algorithms presented here have been implemented in Common Lisp and, more recently, in C++, using the Loom knowledge representation framework (MacGregor (1991)) to maintain the common ground and background knowledge of the hotel application domain. Several components, for example, the *match&substitute* and *add_domain_restriction* functions, have not yet been implemented in a fully general way, and currently handle only simplified cases. The example dialogues discussed in the next section demonstrate how the resolution procedure works.

[17]The current formulation, of first parsing, then determining the contextually understood meaning, rules out interactions that might be desirable. Here we make a trade-off, opting for more straightforward software engineering over the possible benefits that could be derived from using information as soon as it is available (e.g., using contextual information to help in the parsing step). Of course, in taking this approach, we are not making any claims about the way humans do processing.

[18]Note that this is a simplification of what actually occurs in the implementation, where what gets asserted to the CG is not a logical form containing operators, but rather an assertion where knowledge-base instances have been appropriately substituted for variables, and operators have been removed, according to the resolution algorithms.

```
resolve_term(Term)
Let OP(VAR, RESTR, NS) = Term
%%% this representation uses destructuring by pattern matching, as may be
%%% familiar from Prolog, where OP is the top-level operator of Term, VAR
%%% the top-level variable, RESTR the top-level restriction, and NS the top-
%%% level nuclear scope.  This shorthand is used in figures which follow.

%%% Process embedded formulas inside-out
(CULF_R,  CDRS_R)  = determine_CULF(RESTR)
(CULF_NS, CDRS_NS) = determine_CULF(NS)

if OP is a non-presuppositional operator:
   DomRestr = add_domain_restriction(VAR, CULF_R, QUD)
   return (OP[VAR,DomRestr,CULF_NS], CDRS_R ∪ CDRS_NS)
```

Figure 2: Generic resolution algorithm for non-presuppositional operators

4 Discussion of the Example Dialogues

In this section, we discuss the dialogues given in the introduction (repeated here), and highlight how the presupposition resolution operators and algorithms can be used to resolve pronouns, definites, and quantifiers in general (i.e., reference related presuppositions, under our view) as well as other presuppositional phenomena, such as elliptical questions.[19] We illustrate the crucial changes which take place to the QUD data structures, allowing effective resolution of referents and presuppositions.

While the Utterance LF (ULF) describes only the literal content of an utterance, the CULF, along with the CDRS, can be thought of as a record of what the utterance really means, in the context in which it is said. For example, the following ⟨ULF, CULF, CDRS⟩ triple illustrates the QDL structure that results from question (2) of (Dialogue 2) (*What dates will the reservation be for?*):

(QDL 1) $\langle \lambda[x, date(x), def[y, reservation(y), for\_time(y,x)]],$
$\lambda[x, date(x), def[y, reservation(y) \wedge$
$\exists[z, hotel(z) \wedge near(z, MSG), at\_loc(y,z)],\ for\_time(y,x)]],$[20]
{(x:date ?)(y:reservation ?)}⟩

[19] The careful reader will note that these dialogues contain additional reference resolution problems, such as *one*-anaphora (Dialogue 2) and a nonplural antecedent for *they* (Dialogue 3), etc., not discussed here for brevity.

[20] This CULF corresponds to what would arise given one of the possible parses for sentence 1 of (Dialogue 2) (note that there is no conjunct corresponding to the user wanting to *make y*; there are a number of subtle issues in instances such as this which still need to be worked out in this research, in terms of exactly what information is available, e.g., as mentioned earlier, *add_domain_restriction*

Each discourse referent in the set of CDRS is shown in the form *(variable:type instance)*, where *variable* is a logical variable from the CULF, and *instance* is an object in the model (i.e., it is from the knowledge base which represents the common ground–thus the CDRS can be thought of as a list of variables and their bindings). One fact to keep in mind when viewing the dialogues is that questions always produce a new QDL on top of the QUD stack, and therefore a new CULF and CDRS, while answers may update the CDRS of the QDL on top of the QUD stack (i.e., answers may introduce new entities, but these will be added to the CDRS of the relevant question), but answers never produce a new QDL. Another important point is that in the CDRS, there are discourse referents only for what is actually said in the given utterance–thus there is no referent for *hotel* in the CDRS in the example above. This is consistent with the idea that we need contextually enriched information for presupposition resolution, but do not want to create additional entities which could be referred to, for example, in pronoun resolution.[21]

4.1 Pronominal Anaphora: (Dialogue 1)

We will focus on the resolution of the pronoun *it* in the final utterance (8). We claim that at any time there is a set of accessible entities in the discourse, and when a pronoun is used in a discourse felicitously (i.e., as constrained by Relevance), there needs to be a unique maximally salient discourse referent for the pronoun belonging to this set of accessible entities. Under our approach, the set of accessible entities is represented by the union of the CDRSs of all entries on the QUD stack (again, entities mentioned in non-questions are also (potentially) available, since they are included in the CDRSs of the relevant questions on the QUD). Salience is a partial ordering on this set determined primarily by two factors. First, the members of the CDRS of each entry on the QUD stack are more salient than those for all entries below it on the stack. Second, the relative salience of discourse referents within the CDRS of a single QDL is determined by local constraints, such as those given by centering theory (cf. Grosz, et.al. (1995)), or the theory of focusing developed by Suri and McCoy (1994). Our overall approach could be adapted to use any theory of local coherence to determine a partial ordering over the CDRS within a discourse segment corresponding to a single QDL, but it is similar to Suri and McCoy's approach in allowing the CDRSs

still needs to be refined). Different earlier parses would thus lead to different contexts, affecting CULFs for utterances which follow–but for any given system/user interaction we will necessarily only choose one parse.

[21] In other words, we create an entity for *hotel* when that word is actually used, and that entity may or may not be available later to be referred back to. However, when we add conjuncts, for example, to the restriction for y in the example regarding the hotel, z, we don't want to make the mistake of reintroducing a referent that is already present in the discourse, since this could have negative effects later on, for example in the determination of salience, and the like. Again, we are not claiming that these entities are not available for reference, rather, the point is they were already made available (otherwise, we would never have been able to retrieve them in the first place) in the appropriate place in the discourse where they were used.

```
resolve_term(Term)
Let OP(VAR, RESTR, NS) = Term

%%% Process embedded formulas inside-out
(CULF_R,  CDRS_R)  = determine_CULF(RESTR)
(CULF_NS, CDRS_NS) = determine_CULF(NS)

RANKED_REFERENTS = rank_accessible_referents(QUD, CULF_R)
REF_SET = maximal_elements(RANKED_REFERENTS)
If singleton(REF_SET),
     %%% assume REF_SET = {INST}, substitute INST for VAR in CULF_NS
     return(CULF_NS[VAR->INST], {(VAR REF_SET)} ∪ CDRS_NS)
else report no salient referents or failure of uniqueness presupposition
```

Figure 3: Resolution algorithm for the pronoun operator

of prior questions to be stacked. Further explanation of how centering constraints can be integrated with our approach is given by Roberts (1998). In our implementation of pronoun resolution (see Figure 3), the function *rank_accessible_referents* gives the partial ordering of the accessible entities from the QUD, filtering out all entities that are incompatible with the agreement features of the pronoun,[22] which are represented in the restriction component of a pronoun term.

(Dialogue 1)
1) USER: I'm looking for a hotel for June 15th in New York.
2) SYS: What part of the city would you prefer?
3) USER: Manhattan, near Central Park.
4) SYS: How many nights?
5) USER: Just 1.
6) SYS: Will anyone be traveling with you?
7) USER: No.
8) USER: Oh, I want **it** to have a swimming pool too.

In processing this dialogue, the system treats sentence (1) as a question (requests and statements of need and desire should be coerced to questions), and produces (CDRS 1), which is the set of discourse referents mentioned in sentence (1).

(CDRS 1) {(x:person user)(y:hotel ?)(z:date D1)(w:city NYC)}

[22]In reality, things aren't quite this simple, as pointed out by Carl Pollard (p.c.), since agreement features of pronouns and their antecedents do not always match, as in the number mismatch in the following example (as well as in the previously mentioned similar problem in (Dialogue 3)):
Are you sure you checked **every hotel***?*
Yes, **they** *are all full.*

As the system attempts to find out more specific information (imagine that it is filling out a template), it asks subquestions, such as (2), (4), and (6). After each subquestion, a new entry is added on top of the QUD stack, and therefore a new CDRS as well, e.g., the set of discourse referents in the top QUD entry after (2) is (CDRS 2).

(CDRS 2) {(w:city NYC)(x:area ?)(y:person user)}

When a subquestion is answered, as in (3), the CDRS of the current QUD is updated, e.g., the referent (x:area ?) becomes (x:area Manhattan), and a new referent introduced in the answer is added: (z:area CentralPark). However, once a question is completely answered it is popped off the stack. Thus, after (3) is completely processed as an answer to (2), the stack is popped, and later subquestions are also popped after processing (5) and (7). Therefore, when we arrive at (8), the QUD stack is just as it was after (1), since all of the intervening subquestions have been popped. This approach accounts for the observation that more recently mentioned entities, such as *Manhattan* or *Central Park*, are less likely as antecedents for *it* than those from (CDRS 1), which are closer in terms of hierarchical discourse structure.

In order to determine the antecedent for *it*, *rank_accessible_referents* only has to consider (CDRS 1), returning a subset from which (x:person user) is removed, because a person, being animate, does not match the restrictions of *it*. Thus, the search for possible antecedents has been significantly constrained by using the CDRS associated with the QUD. Among the remaining elements, the most likely antecedent is (y:hotel ?), which we call an *unanchored discourse referent*, since it is not yet bound to an actual instance of a hotel. This might be ranked highest by some versions of centering theory, because it is a complement of the verb, while the other referents were introduced by adjunct phrases (*for June 15th* and *in New York*). In general, however, pragmatic plausibility must be considered as an additional filter when determining whether a candidate is a potential antecedent. For example, (z:date D1) can be ruled out because it is not plausible for dates to have swimming pools.

4.2 Definite Descriptions: Dialogues 2–4

Although definite descriptions can often be identified with antecedents from the CDRS in essentially the same way as pronouns (since each CDRS is a subset of the CG Domain), they are not required to corefer with a maximally salient discourse referent. Therefore, our algorithm specifies three ways for a definite reference to be resolved. First, we check whether the CDRS accessible on the QUD stack contains a unique element that matches the restriction of the definite operator. Second, if there is no salient antecedent of the appropriate type, then we attempt to find a unique entity in the CG which satisfies the restriction. Third, if this fails, we use accommodation where possible to introduce an entity from the application's database into the

```
resolve_term(Term)
Let OP(VAR, RESTR, NS) = Term

%%% Process embedded formulas inside-out
(CULF_R,  CDRS_R)  = determine_CULF(RESTR)
(CULF_NS, CDRS_NS) = determine_CULF(NS)

REF_SET = all_accessible_referents(QUD, CULF_R) % possible anaphoric reference
If singleton(REF_SET),
   return(CULF_NS[VAR->INST], {(VAR REF_SET)} ∪ CDRS_NS)
else if |REF_SET| > 1,
   report failure of uniqueness presupposition
else % no salient antecedent, retrieve referent from common ground
   DomRestr = add_domain_restriction(VAR, CULF_R, QUD)
   REF_SET = retrieve_referents(VAR, DomRestr, CG)
   If singleton(REF_SET),
      return(OP[VAR,DomRestr,CULF_NS], {(VAR REF_SET)} ∪ CDRS_R ∪ CDRS_NS)
   else if |REF_SET| > 1,
      report failure of uniqueness presupposition
   else % attempt to accommodate, retrieve referent from application database
      REF_SET = retrieve_referents(VAR, DomRestr, ApplicationDB)
      If singleton(REF_SET),
        return(OP[VAR,DomRestr,CULF_NS], {(VAR REF_SET)} ∪ CDRS_R ∪ CDRS_NS)
      else if |REF_SET| > 1,
        report failure of uniqueness presupposition
      else report failure to accommodate
```

Figure 4: Resolution algorithm for the definite description operator

CG (see Figure 4).

(Dialogue 2)	1)	USER:	I want to make a reservation at a hotel close to Madison Square Garden.
	2)	SYS:	What dates will the reservation be for?
	3)	USER:	March 3rd and 4th.
	4)	SYS:	Would you like a single room?
	5)	USER:	Yes.
	6)	USER:	Also, I'll need a conference room on the 4th.
	7)	USER:	I'd prefer it if **the hotel** had one.

In (Dialogue 2), we focus on the resolution of *the hotel* in sentence (7). We first look for an appropriate antecedent in the CDRS accessible on the QUD stack, as in our treatment of pronominal anaphora, so we need to trace the stack for this

dialogue. A request is made by the user in sentence (1), followed by a series of specific questions generated by the system. The QUD after (1) has the following CDRS:

(CDRS 1) {(x:person user) (y:reservation ?) (z:hotel ?) (w:place MSG)}

Subquestions are asked in (2) and (4) ((2) and (4) are subquestions of (1), which is treated as a question in the same way as sentence (1) of (Dialogue 1) was earlier, both being coerced to questions since they are statements of need or desire) and answered in (3) and (5), respectively, so the QUD stack is pushed and popped, but at (6), it is at the same state as it was after (1). (6) is interpreted as a request, so a new entry with (CDRS 6) is pushed onto the QUD on top of the QDL for (1).

(CDRS 6) {(x:person user) (v:conf-room ?) (u:date D4)}

In order to interpret the definite description anaphorically, we search for discourse referents whose type satisfies the explicit *hotel* restriction within the set of all accessible CDRS, viz., the union of CDRS 6 and CDRS 1. Since this set contains exactly one referent (z) which matches the *hotel* type, the uniqueness presupposition is satisfied and z is selected from CDRS 1 as the antecedent.

It is also possible for a definite description to have no explicit antecedent, as in *the Marriott*[23] in sentence (3) of (Dialogue 4). In such cases, an empty set of referents will be returned by *all_accessible_referents*, and our algorithm will attempt to retrieve a referent from the common ground. Before resolution, the content of this description is DEF 3, in which the variable ?NS is a placeholder for the unspecified nuclear scope of the def operator:

(DEF 3) $def[y, Hotel(y) \wedge Named(y, Marriott), ?\text{NS}]$

The restriction of this term is obtained from the lexical entry for *Marriott*, which contains the information that it refers to a hotel, in addition to specifying its name. Although we rely on domain-specific knowledge in assuming that it refers to a hotel, we believe this assumption is reasonable, because the proper names for hotels can be automatically acquired from the hotel database used by the application.[24]

Now suppose that there are a number of Marriotts in the area. In an empty discourse context, this reference would have an unsatisfied uniqueness presupposition, so the system would need to ask the user which Marriott was intended. However, in

[23] Some might interpret the use of *the Marriott* in this dialogue as more of a name (i.e., the hotel chain as a whole, rather than a specific hotel near the airport), than a definite description. This is another difficulty that a system will have to cope with.

[24] We do not want to claim that linguistically this is the proper treatment in general, rather, it is a feature of having a fixed domain for the dialogue that we can take advantage of.

this case, uniqueness can be established by searching the QUD for an appropriate domain restriction, which can be conjoined with the explicit restriction given in (DEF 3). Since domain restrictions can be contextually supplied for most restricted operators, we interpret (DEF 3) as if there were an additional conjunct, which is schematically represented by QUD_RESTR(x) in (DEF 3′):[25]

(DEF 3′) $def[y, (Hotel(y) \wedge Named(y, Marriott) \wedge \text{QUD\_RESTR}(x)), ?\text{NS}]$

As in our treatment of anaphora, the key to constraining the search for an appropriate domain restriction is the QUD structure of the discourse. The entry on top of the QUD corresponds to question (1) of (Dialogue 4), whose CULF is (simplified):

(CULF 1) $\lambda[x, Hotel(x) \wedge Near(x, Airport), \exists[y, Date(y), HasVacancyOn(x, y)]]$

To determine whether any implicit domain restriction can be added to *the Marriott*, our algorithm calls *add_domain_restriction* to search the QUD for predicates that match the same basic type as the explicit restriction, *Hotel*. In (CULF 1) it finds the restriction $Hotel(x) \wedge Near(x, Airport)$,[26] which can be added in place of the virtual QUD_RESTR(x) conjunct in (DEF 3′) to further restrict the domain for *the Marriott*. This restriction (DEF 3″) is then used by *retrieve_referents* to find a matching referent in the CG.

(DEF 3″) $def[y, Hotel(y) \wedge Named(y, Marriott) \wedge Near(y, Airport), ?\text{NS}]$

It is important to note that the familiarity presupposition for a definite description does not require its referent to be previously mentioned in the discourse. In sentence (1) of (Dialogue 3), the referent for *the Holiday Inn* does not yet exist in our representation of the common ground, because the system initially has no knowledge that the user is aware of any particular Holiday Inns. In such cases, no objects are

[25] We do not actually include an explicit conjunct for the domain restriction in our implemented logical forms, because an implicit domain restriction may be added to virtually any restricted operator, as motivated by Roberts (1995), and it is of course possible for no new information to be added by domain restriction.

[26] Note that in this example, the user has no way of knowing if the system is using a *mention all* or a *mention some* strategy in its answer in (2). Thus (3) could take several meanings. The one assumed here (which seems the most likely): The user wants to check whether there is a Marriott at the airport with vacancies. Another potential interpretation might be: Being aware of an airport Marriott, the user wants to make sure that it too is full, before moving on. A third and perhaps less likely interpretation is: The user wants a Marriott at all costs, regardless of location near the airport. In instances such as these, the system will err on the side of assuming that elliptical questions relate as straightforwardly as possible to previous questions, since the sort of inferencing to determine the right interpretation is computationally expensive, and it is debatable whether a single interpretation is uniformly preferred by all speakers.

```
resolve_term(Term)
Let OP(VAR, RESTR, NS) = Term

%%% Process embedded formulas inside-out
(CULF_R,  CDRS_R)  = determine_CULF(RESTR)
(CULF_NS, CDRS_NS) = determine_CULF(NS)

DomRestr = add_domain_restriction(VAR, CULF_R, QUD)
case non-top-level OR non-question:  % non-presuppositional
    return (OP[VAR,DomRestr,CULF_NS], CDRS_R ∪ CDRS_NS)
case wh-question:  % presupposes some object satisfies DomRestr
    REF_SET  = retrieve_referents(VAR, DomRestr, CG)
    return (OP[VAR,DomRestr,CULF_NS], {(Var REF_SET)} ∪ CDRS_R ∪ CDRS_NS)
case polar-question:
    return (OP[VAR,DomRestr,CULF_NS], CDRS_NS)
```

Figure 5: Resolution algorithm for the lambda operator

returned from the CG by *retrieve_referents*, and the *definite* presuppositional term will remain with an unknown referent in the output of *determine_CULF*. Our approach to accommodation for such unsatisfied presuppositions (see Figure 4) is to look for a referent in the application's private database of facts about the domain of hotels, since this database represents all of the world knowledge that the system has available. Thus, the application must make its database readable to the dialogue system (i.e., it must provide an interface for read-access only). If the dialogue system finds a unique hotel named Holiday Inn, we can assume this hotel satisfies the user's presupposition. On the other hand, if it turns out that there are either no hotels named Holiday Inn in the application database, or multiple Holiday Inns, the system could report the failure of these presuppositions, rather than giving an uninformative simple negative answer to the user's question (1).

4.3 Generalized Domain Restriction: (Dialogue 3)

Consider next the quantificational determiner *every* in sentence (3) of (Dialogue 3). It should be clear that the user is not asking about every morning for all time, but only about all mornings during the planned trip. As with definite descriptions, our algorithm allows the restriction of most operators with semantically contentful restrictions[27] to be further specified by information from the QUD, so the

[27]Domain restriction is not usually applicable to pronouns and other expressions that have little explicit content, because these expressions depend on recovering a salient antecedent in order to determine the type of the referent, rather than searching for a particular type of object in the common ground.

interpretation of *every morning* will differ depending on whether the dialogue began with question (1a) or (1b). Now, if it is the case that the Holiday Inn has a breakfast buffet on weekdays only, it is important for the system to answer (3) appropriately, as in (4a) and (4b), depending upon the context created by (1a) and (1b).

(Dialogue 3)
1) USER: Does **the Holiday Inn** have any vacancies for
   a) Tuesday, 12/4 - Friday 12/7?
   b) Thursday, 12/6 - Saturday 12/8?
2) SYS: Yes, several.
3) USER: Do they have a breakfast buffet **every morning**?
4) SYS: a) Yes, Monday through Friday.
   b) No. There's a breakfast buffet Monday through Friday, but none on Saturday.

To determine the domain restriction for *every morning*, *add_domain_restriction* searches the QUD for predicates that match the same basic type as the explicit restriction, *morning*. In this case, we take the basic type to be a temporal entity, so it will search for temporal descriptions in the QUD.[28] By using the QUD stack to constrain the search, *every* will quantify over any temporal entities that are found at a level of discourse structure closest to the current segment, but crucially not over every temporal entity in the entire common ground. Thus, to determine the response in (4a), only the date range mentioned in (1a) is relevant, and a positive response can be given, since the question relates to weekdays. In (4b) however, the date range includes a Saturday, so the system should generate a negative response. Thus, the system has employed domain restriction in answering *Yes* in (4a) and *No* in (4b), but note that in both cases that the system reports also to the user that the breakfast buffet is available *Monday through Friday*, even though these days do not exactly match the date ranges provided by the user. The reason for this is simple: In attempting to be as cooperative as possible, the system tries to provide complete information where it can.[29] In this manner, the user is not misled (that there is no buffet available on Monday), for example, had the system also used only the domain restricted dates in (4a), and answered: *Yes, Tuesday through Friday.* Here, as elsewhere, there are a number of tradeoffs in what sorts of strategies yield the most cooperative system, and clearly these could benefit from empirical evaluation of human-to-human systems, as well as human response to the given system, once it is complete.

[28] A complete explanation of this situation might require the system to infer the domain goals of the user. However, when the QUD contains some descriptions of the appropriate type, we can use them as an approximate domain restriction, thereby avoiding the computational expense of full plan inference (see section 5).

[29] But, of course, this does not extend to situations where complete answers would be too long to be useful, i.e., the *mention some* vs. *mention all* distinction previously mentioned, for example in answering the question: *which hotels have vacancies*, where a *mention all* interpretation would be rather uncooperative if thousands of hotels had vacancies.

```
resolve_term(Term)
  % Assume nuclear scope of Term is of the form: φ[OldExpr->NewExpr]
  while QUD stack is not empty {
     QUD-CULF = CULF of QUD[top]
     QUD-CDRS = CDRS of QUD[top]
     (NewExpr, CDRS1)  = determine_CULF(NewExpr)
     if NewExpr is a generalized quantifier,
        let SubstLF = match&substitute(QUD-CULF,restriction(NewExpr),NewExpr)
     else (NewExpr is a predicate)
        let SubstLF = match&substitute(QUD-CULF,NewExpr,NewExpr)
     if null(SubstLF)
        or SubstLF is not interpretable as a subquestion of QUD-CULF,
           pop(QUD)
     else return (SubstLF, priority_union(CDRS1, QUD-CDRS))
     % priority-union(X,Y) is like set union, but when some members of X and
     % Y have the same type, only the member of X is included in the result.
     }
```

Figure 6: Resolution algorithm for WH_Ellipsis

4.4 Elliptical Questions: (Dialogue 4)

(Dialogue 4)	1)	USER:	Which hotels near the airport have vacancies?
	2)	SYS:	The Holiday Inn and Sheraton have vacancies.
	3)	USER:	How about **the Marriott?**
	4)	SYS:	No, the airport Marriott doesn't have any vacancies.

(Dialogue 4) is a somewhat more complex dialogue, including an elliptical question as well as several definite descriptions. It illustrates how our approach generalizes to the larger class of presuppositional constructions which we identified in the introduction. Let us focus on the interpretation of sentence (3), *How about the Marriott?*, which is assigned the following ULF:

(ULF 3) $Wh\_Ellipsis[\phi, Question(\phi),$
$\phi[X \rightarrow (def[y, Hotel(y) \wedge Named(y, Marriott), ?\text{NS}])]]$

ϕ is a variable referring to some contextually salient question, and the definite description corresponding to *the Marriott* is to be substituted for some term (X) within ϕ. Recall that the variable ?NS is a placeholder for the unspecified nuclear scope of the def operator.

The presuppositional operators process the logical form of an utterance *inside-out*, i.e., the embedded context resolution problems are handled first, so first the def

term corresponding to *the Marriott* is resolved, as we discussed in §4.2 on definite descriptions, and *add_domain_restriction* produces the refined description (DEF 3″):

(DEF 3″) $def[y, Hotel(y) \wedge Named(y, Marriott) \wedge Near(y, Airport), ?\text{NS}]$

Next, the top-level *Wh_ellipsis* term in (ULF 3) is resolved, according to the *resolve_WH_Ellipsis* algorithm of Figure 6. ϕ must be a question, so we retrieve the question on top of the QUD stack, and attempt to identify ϕ with its CULF (CULF 1).

(CULF 1) $\lambda[x, Hotel(x) \wedge Near(x, Airport), \exists[y, Date(y), HasVacancyOn(x, y)]]$

We must now find a term within (CULF 1) for which the term corresponding to *the Marriott* can be substituted. Our *match&substitute* algorithm looks for terms whose restrictions specialize a common basic type, so it again finds the restriction on the (top-level) λ-term containing the *Hotel* predicate in (CULF 1).

The operator and restriction of this term (i.e., (CULF 1)) are replaced by those from (DEF 3″) and the variables are unified, but the nuclear scope of (DEF 3″) is unspecified, so the nuclear scope of (CULF 1) remains unchanged in the result:

(3″) $def[x, Hotel(x) \wedge Named(y, Marriott) \wedge Near(x, Airport),$
$\exists[y, Date(y), HasVacancyOn(x, y)]]$

(3″) is (almost) the CULF for *How about the Marriott?*, but it must be noted that it should be interpreted as a polar question, since the λ-term characteristic of a wh-question has been replaced by a definite description.[30]

Thus, both the elliptical question and the domain restriction of the definite description are processed by the same overall strategy: They are interpreted by incorporating information contained in the question under discussion.

5 SDRT and its Relation to the Present Theory

In the preceding sections of this paper, we presented an architecture for a dialogue system, and described how it would handle several example dialogues. But, of course, in the research on discourse, a number of such theoretical approaches have been developed. In this section, we briefly describe one widely-used system, Segmented Discourse Representation Theory (SDRT), and then contrast it with our own, especially in terms of what we estimate to be the computational resources required for each approach.

[30]When all top-level λ-terms in a wh-question have been replaced, it is interpreted as a polar question.

5.1 The Roots of SDRT

SDRT, as its name implies, is a descendant of DRT (for a comprehensive introduction to DRT, see Kamp & Reyle (1993)). DRT is a semantic theory which focuses on the representation of discourse, and may be thought of as a dynamic, procedural theory, reflecting the idea that sentences are interpreted step-by-step, as they are uttered. As such, it presents a radical departure from Montague grammar, in both its reliance on the procedural nature of a discourse, and its focus on series of sentences rather than a single sentence.

Like Montague grammar, DRT offers a logical representation of utterances, but diverges in where and how the logical representations are stored. In Montague grammar, syntactic structures are translated directly into logical structures, and these logical structures may then be interpreted with respect to a model. It is in theory possible to eliminate the translation step, since the logical structures derived are a function solely of the parts from which they were formed (i.e., the whole is no greater or less than the sum of its parts). In DRT, however, there is also a representation of the context in which a sentence occurs. Thus DRT has an additional level of representation, and it is in this level that logical representations of sentences are stored (this will become clearer below). Because of these differing levels, interpretation in DRT proceeds rather differently. Rather than interpreting a sentence directly with respect to a model, in DRT the truth of a discourse is defined in terms of whether the information contained in the representation of the discourse can be embedded in a model.[31] It is in this sense that many view DRT as noncompositional, since the interpretation of an utterance is not a function of just the parts of the utterance, but rather a function of the utterance and the previous context, yielding an updated context. This view is not really correct, however, since there are also compositional versions of DRT (for example, Zeevat (1989)).

The development of DRT was partially motivated by phenomena that were problematic for Montague grammar, such as intersentential anaphora. For example, in the discourse:

(5) Exactly one person made a reservation. She's staying a week.

A Montagovian system cannot handle this type of discourse, because the antecedent for the pronoun *She* occurs across a sentence boundary. And the solution of conjoining the logical representations of the two sentences does not work either, because the process of quantifying in the term *Exactly one person* after the sentences are conjoined creates the wrong meaning, i.e., we get:[32]

(6) $\exists y \forall x((Person(x) \wedge Made\_a\_reservation(x) \wedge Stay\_a\_week(x)) \iff x = y)$

[31] Much of the material in this section relies heavily on chapter 7 of Gamut (1991).

[32] In this section, we avoid the generalized quantifier notation used elsewhere, simply to maintain a more transparent relation to the Montague style.

where some other person could have made a reservation, rather than what we want:

(7) $\exists y \forall x((Person(x) \wedge Made\_a\_reservation(x) \iff x = y) \wedge Stay\_a\_week(y))$

DRT resolves a number of such difficulties. In DRT, the variables are called *markers*, and the formulas in which take markers as arguments are called *conditions*. Markers are introduced for indefinites and proper names, and are a way to keep track of the individuals mentioned in a discourse–the *discourse referents*. There are two common and equivalent notations for DRT, a linear notation, and DRT's distinctive "box" notation. We will use the latter here. The general idea is that the context of a discourse is represented by the box, and in the box are the markers and the conditions. Thus, the boxes help determine the scope (DRT theorists tend to refer to the *accessibility* of a marker) of the markers, and the graphic representation provides an intuitive feel for exactly what is accessible when. For example, in the one sentence discourse:

(8) Mary visits a hotel.

There will be markers, x and y, for the two discourse referents mentioned (*Mary* and *a hotel*) as is shown in the following box, which is called a DRS[33] (Discourse Representation Structure), where markers are at the top, and conditions below:

(9)
```
| x y |
|-----|
| Mary = x |
| hotel(y) |
| x visits y |
```

We will not go into the details here of how such DRSs are constructed, nor the rules for accessibility or interpretation, rather, at this stage we are trying to give the reader a basic feel for how DRT works, so that we can give a similar sketch of SDRT. Continuing on then, a DRS for the discourse in (5), would look as follows:

(10)
```
| x |
|---|
| Person(x)                                  |
| Made_a_reservation(x)                      |
| Stay_a_week(x)                             |
| | y |                                      |
| | --- |                                    |
| | Person(y)                              |
| | Made_a_reservation(y)                  |
| | ---------------------------------------|
| | → | | |
| | | x = y |
| | | ------ |
```

[33]This should not be be conflated with the CDRS, mentioned in earlier sections of the paper.

The major point to take from here is that the person referred to in the first sentence, is available to be referred back to (i.e., it is accessible), illustrated graphically by the fact that all the conditions on x are inside (however deeply embedded) the same box where x is introduced as a marker.[34]

5.2 SDRT

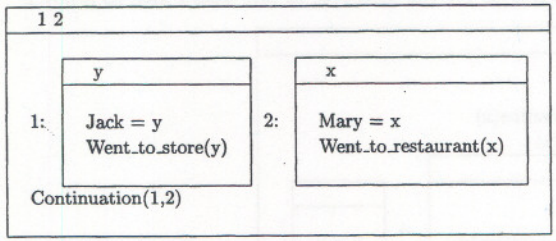
SDRT (Asher (1993), Lascarides & Asher (1993)) builds on DRT by adding the notions that DRSs should be related to one another in a more formal way (i.e., rather than simply being contained in or merged into, for example, the same larger DRS). SDRT employs rhetorical relations, along the lines of Mann & Thompson (1987), to relate DRSs. These rhetorical relations, the number and types of which are often debated in the literature, include relations such as *elaboration*, *explanation*, *background*, and *contrast*, and are an attempt to describe the way one unit of a discourse (often a sentence) relates to another (i.e., what pragmatic purpose does it serve). Thus, SDRT is a theory of the structure and content of discourse.

SDRSs are like DRSs, but each SDRS has a unique label, and in addition to containing DRSs and other SDRSs, an SDRS can contain conditions where rhetorical relations are the predicates and labels of SDRSs are the arguments. In short then, in SDRT we see a new type of marker, i.e., the labels representing SDRSs, and a new type of condition, but (at least graphically) things are very much the same as in DRT. For example, in the discourse:[35]

(11) Jack went to the store. Then Mary went to the restaurant.

The SDRS constructed might look as follows, where we assume the rhetorical relation is *Continuation*:[36]

(12)



[34] Things in DRT are obviously much more complex than this, e.g., discourse markers are accessible when DRSs are *subordinate* to each other, for example when connected by →, the markers of the *antecedent* (i.e., left) DRS are available to the *consequent* (i.e., right) DRS.

[35] We use a simple example here, to avoid issues with quantifiers, anaphora, etc.

[36] Again, this is a simplification of what the theory would yield.

By relating SDRSs with rhetorical relations, SDRT represents the hierarchical structure of the discourse, and captures important meaning differences between discourses such as (13) and (14):

(13) Jack smiled. Then Mary told her joke.

(14) Jack smiled. Mary had told her joke.

where in the first case the rhetorical relation would likely be *Continuation* and in the second case *Explanation*.

In providing this hierarchical and discourse content, SDRT provides quite a bit of information which can be used in difficult areas, such as the presupposition and anaphor resolution discussed earlier in this paper. But the additional information does come at a price, compared to DRT. Whereas in DRT, an update function (i.e., a function to update the global discourse context after a new utterance has occurred), can be as simple as unioning the markers and conditions of the new utterance with those of the discourse so far, in SDRT the update function becomes much more complex (see Asher & Lascarides (1998b) for a detailed discussion). In addition to computing which rhetorical relation to use, the attachment point for the new information must also be computed. Asher and Lascarides accomplish these steps with a nonmonotonic logic, and a number of constraints on the types of discourses which can occur. For example, much like Gricean maxims, attachments and relations which maximize discourse coherence are preferred. By constraining the SDRS construction in this way, Asher and Lascarides make strong claims about which information is available at later points in a discourse, and they provide a mechanism for adjusting the discourse structure which can be adapted depending on the current theory, or for that matter the language or style of discourse.

In addition to the aspects already mentioned, SDRT extends DRT by attempting to model the intentional structure of the participants in a discourse. This can have advantages, in terms of discourse processing, since as a number of works have shown (e.g., Grosz & Sidner (1986)), modeling dialogue can require knowledge of the plans of the participants. But, of course, this addition also comes at a computational cost, as will be discussed shortly.

Extending SDRT to dialogue is straightforward, and indeed seems a natural fit. In Asher & Lascarides (1998a), new rhetorical relations are introduced which are appropriate for questions in dialogue, including Question Answer Pair (QAP), Indirect Question Answer Pair (IQAP), and Question Partial Answer Pair (QPAP), along with new axioms on when the relations can be used.[37] The (computed) intentional structure of the participants plays a key role in using these new relations. Because

[37]Here, as elsewhere, for brevity we do not go into the inner workings of the SDRT approach, but again try to provide an overview of the benefits and costs.

of this, Asher and Lascarides must maintain separate SDRSs for each discourse participant, since not only may different participants' intentions vary, but so may their perceptions of intentions (e.g., x believes y intends z, and y knows x believes y intends z, and x thinks y doesn't know x believes y intends z, and so on).

5.3 Comparing the Current Approach with SDRT

We are now at a good point to view the simplified architecture presented earlier in this paper in SDRT terms, and to compare the two approaches. It should be emphasized that we are not attempting to point out weaknesses in SDRT here. Rather, we hope to demonstrate that in certain contexts, such as query-based human/computer dialogue where the domain goals are fixed, that a simplified architecture will often be sufficient, and therefore efficiency gains come at a low cost in terms of coverage. Indeed, in many ways, the simplified approach is a subset of SDRT. Further, our approach is a work in progress, since the entire system is not yet implemented,[38] and we hope that as the work in SDRT advances, it will help inform our own.

First, one of the obvious differences between SDRT and our approach is that we maintain uniform discourse structures, rather than separate discourse structures for each participant.[39] We are able to make this step because we do not (at least not in the discourse structures, see the next section) explicitly model the intentional structure of the participants. Indeed, when the domain goal(s) are fixed, as in the case of a hotel reservation system, the intentions of the human participant can be assumed (i.e., the overarching goal of making a hotel reservation). From the perspective of SDRT, our approach could be viewed as one where all participants happen to have the same SDRSs, i.e., they agree on their perceptions of each other. This brings up an important theoretical point. If interlocuters are engaging in cooperative dialogue, then they are following the rules of the dialogue game, i.e., they are obeying Gricean maxims, and the like. Therefore, we might want to require that there be uniform discourse structures, such as the CG. Discourse is a communal activity, and for it to be successful, the participants have to agree on what the goals and state of the discourse are.[40] From this point of view, the separate discourse structures of SDRT do not appear to be highly motivated. One final point regarding the issue of uniform vs. separate discourse structures is that, in general, maintaining single, uniform structures for the discourse, rather than for each participant, should reduce both the space and time involved in computation.

[38] Consequently the efficiency gains are necessarily only an estimate. Plus, we are not aware of any completed implementations of SDRT, although a promising demo was given by Bohlin & Larsson (1999). So, the comparison in this section is a bit difficult from the get-go, since we compare our theory and partial implementation with a (at this point) more explicit theory and no implementation; nevertheless, some important differences do arise.

[39] Another type of QUD structure by Ginzburg (1996) maintains a separate QUD for each participant, but we do not believe this to be necessary for this type of dialogue.

[40] Thanks to Craige Roberts for pointing this out to me.

A more important difference centers on our reliance on the QUD stack. The QUD functions as both the representation of the discourse hierarchy and as the implicit store of some of the rhetorical relations between utterances (i.e., it serves to relate questions and answers, and superquestions with subquestions), while the informational content of the discourse is ultimately stored in the CG. In SDRT, on the other hand, all of this information (as well as the intentional structure) is stored in the SDRSs. This has the advantage of making all the information available in one place for processing, but sacrifices modularity (we return to this in the next section). The implicit relations in the QUD are not unlike those used for questions in SDRT. For example, each question and answer form a question/answer pair, and these relations must be computed under both approaches. However, because we assume that the discourse revolves around the idea of answering the current question under discussion, until the domain goals are completed, we have a much smaller inventory of relations to choose from.[41] We must also keep track of the relation between subquestions and superquestions (again, this is accomplished by using the stack-like data structure, where any question on top of another is a subquestion), and, of course, compute when they occur. Finally, we must also interpret requests and statements of desire as questions. Given this approach, our system does especially well in cooperative dialogues, where users do not switch back and forth between unrelated questions, since such switching might cause the QUD to be popped, and therefore need to be adjusted, once earlier questions are resumed. Thus, our system, as described, does not presently tolerate asides well, but certainly this is not a theoretical limitation of the system. By taking advantage of cue-phrases (such as *by the way*, *that reminds me*, etc.), a system can discern when the question under discussion is being changed to a (possibly) unrelated one, and adjust its discourse structures accordingly.[42] And, indeed, if a dialogue is coherent, interlocuters must adequately signal such changes in the topic of discussion, or they are violating the maxims of cooperative dialogue.

In our approach, the relation between questions and answers is indeed implicit, since we do not store the answer alongside the question (although, importantly, we do store a representation of the discourse referents and information updates contained in answers, in the CULF and CDRS of the relevant question). And, once a question is completely answered, it is removed from the stack. This means, again, that our machinery, like in SDRT, must be able to compute what entails an answer to a question, and must be sensitive to user input to help indicate when questions have been answered successfully (indeed, we rely on this information, and let the user "drive" the dialogue). Note that we do not really lose any information in popping the QUD, since we keep a record of the information content in the CG, as well as a record of the hierarchical structure, since each utterance record, what we call

[41] One of the criticisms of theories that use rhetorical relations is not only that they are difficult to compute, but that those typically used may not be exhaustive enough for many discourses.

[42] An obvious, although inelegant solution would be to simply pop questions onto a second stack-like data structure, where they could be accessed as necessary for re-pushing onto the QUD. Thus, each dialogue would have its own discourse structures as its disposal. We would also need algorithms for backtracking, just as humans do (i.e., *the Okay, now where are we?*).

a Move, points to the QUD at the state it was in when the utterance occurred. By maintaining the QUD (i.e., pushing and popping, as well as updating), we too make claims about which discourse referents are accessible–but unlike SDRT, there's only one possible attachment point, at the top of the QUD. Again, we believe this represents a significant savings in computation.

This now brings us to the question of search space. When we encounter discourse referents which need to be resolved, the first place we look (the strategy differs from operator to operator, but this is the most common approach) is the top of the QUD, where there is a single question, and a number of referents contained in the CDRS corresponding to that question. The search may involve looking at super questions (i.e., lower on the stack) and ultimately, as described in section 4, looking in the CG or possibly being accommodated. We posit that a large proportion of the time, we will find the material in the QUD (the QUD thus functions somewhat analogously to a cache, in terms of finding the most recent or salient referents). Depending on the axioms used in the SDRT implementation, a search may be much slower. However, this appears to be a point where the systems are similar, if the SDRT implementation is sufficiently constrained (e.g., only referents on the right frontier are accessible). But, once again, the computation which must take place beforehand (i.e., what will be attached where) is more expensive in the SDRT case.

In terms of coverage, SDRT appears to be the winner, at least compared to the coverage of our implementation so far. For example, at the moment, our implementation provides no mechanism to refer to entire utterances anaphorically, much less spans of dialogue, or to general ideas/intentions which can be inferred from it. Our hypothesis is that users interacting with a computer will not only tolerate such deficiencies, but quickly adapt to them. Nevertheless, we would be severely mistaken to say that we have complete coverage. But, like other theory vs. implementation concerns raised in this paper, this is not a theoretical limitation, just a mechanism which has not yet been built. Since we already track each utterance (e.g., in the move history, M), and since segments of discourse are represented in the QUD, we could certainly add the means to refer back to utterances or segments of discourse anaphorically (perhaps by some representation or pointer to the appropriate object in the knowledge base). Ideally, our use of the QUD would also help to narrow the search space in searching for the referents–in the case of utterances and discourse segments–since interlocuters would be likely to refer back to the most salient discourse entities (much like earlier our discussion regarding pronouns; for a further look at referring back to 'larger' entities (i.e., whole utterances, propositions, etc.) anaphorically, see Roberts (1995)).

The last, and perhaps most important area in terms of computational cost is the question of modeling intentional structure. We do not, at least not as part of the dialogue system itself, explicitly model the intentional structure of the participants. Thus, as just mentioned, we will miss any references to plans, as well as some resolution problems that rely on the correct identification of intentions. However, given the exceedingly expensive nature of computational inference in calculating plans, we be-

lieve this to be a significant advantage in a domain such as the one described. Again, we are able to take this step because much of the plan structure can be inferred from the domain, and from the roles served by the computer and humans (the computer always helps the human make a reservation, not the other way around). Additionally, we do provide a mechanism for planning to be incorporated by applications (see the next section); we simply do not integrate it into the dialogue system proper.

Thus, an implementation using the architecture we described should be significantly faster than one using SDRT, but the speed-up does come at a coverage cost. And, while in the final section we will describe advantages that a modular approach brings, an integrated approach such as SDRT certainly offers organizational advantages, in that all of the information is in one place (i.e., the intentional structure is included in the SDRSs, not in a separate planning component, and the hierarchical structure is at the same level as the informational content–relations in and between SDRSs–rather than being in separate but linked data structures such as the QUD and CG).

6 Conclusions

In the previous sections of this paper, we have focused on specific parts of the dialogue system, mainly the QUD and CG, and discussed their role in the presupposition resolution process. In this final section, we present the overall architecture, and discuss the benefits of using a modular approach to the dialogue system, and then finally outline our plans for future work.

6.1 Overview of Architecture Described

As shown in Figure 7, the dialogue system is comprised of a number of modules, including the CG, the QUD, a Parser, and a Generator. Each of these modules has a number of operations specified for them, and are only accessible to one another via their interfaces. Thus, as is normally the case with object-oriented design, each part of the system is viewed as a separate entity, with a number of different tasks it can be asked to do. For example, the QUD may be asked to pop itself, the CG to return the answer to a query, and the Parser to parse the utterance spoken by the user. Also shown in the diagram are the operators and algorithms which guide the resolution process, grouped as a unit, although each operator may be viewed as an object (more accurately an instance of an object) in its own right, and the set of conversational moves, the *Move History* (which was abbreviated in earlier sections as M). Once an utterance is superficially parsed, the system checks the output of the parser for any resolution necessary (i.e., for the presence of any presuppositional operators), then asks the operators to evaluate themselves, again by accessing the data structures when necessary, as shown in section 3 earlier. When finished, the now contextually understood logical forms (CULFs) and other information from the sentences are passed to the appropriate data structure modules.

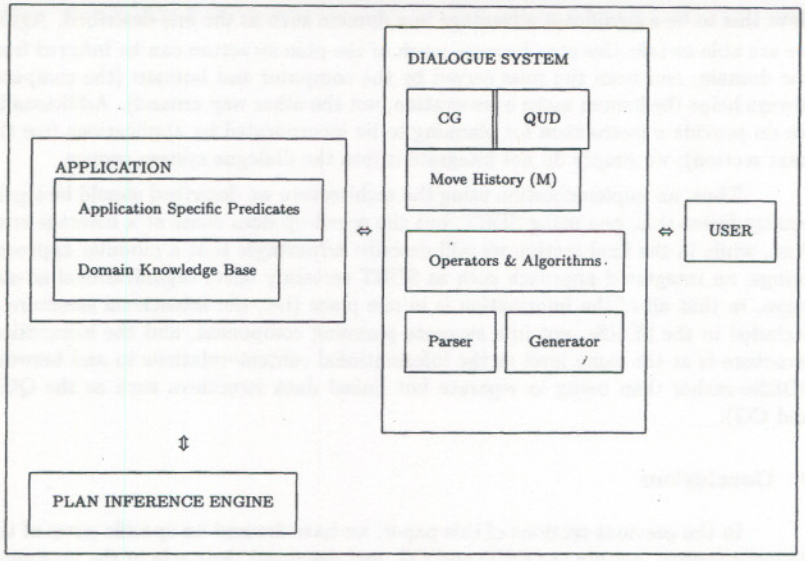


Figure 7: dialogue system architecture

The true modularity of the system comes into play with regard to the application. An application, such as the hotel reservation system described in this paper, may access any of the modules of the dialogue system, and indeed, may inform them. For example, in this instance, the application may have specific predicates which are applicable to the domain, for example *have-vacancies(h,d)* (where h is a hotel and d is a date), that would most likely not be relevant for other domains such as an airline ticket booking system. The dialogue system itself will have been initialized with a number of expected predicates for dealing with queries, monetary transactions, and the like. Application specific predicates may then be made available to the dialogue model (i.e., for this application only) to supplement the terminology in the CG for processing. This is consistent with the idea that most users contacting a hotel reservation system will know how hotels function. Thus the system has enough information to handle generic query-processing dialogue, and can be customized for specific task domains.

The overall interaction of the application and the dialogue system can be seen as a series of messages being passed. The application may receive an indication (say, from a user interface) that the user has communicated something. The application then asks the dialogue system to begin processing the input. The system will, as described, parse and resolve the input, and after updating the various data structures,

signal the application, and optionally pass the logical form to the application.[43] The application may then (if desired) access the data structures, in order to decide what step it will take, if any. In this sense, the dialogue system is a *dumb* system. It stores the information, and then allows the application to retrieve information, to get an accurate view of the state of the discourse at any time. Here is where an optional planning engine might come into play. As mentioned earlier, plan inference is expensive, and may not always be desirable. But, if opted for, an application can pass information from the dialogue system to a planner (see Figure 7), to help it decide what conversational act it should next perform. Again, we see this modularity as an advantage, because it allows an application to be as intelligent as it can computationally afford to be. Certainly, we assume some kind of planning capability must be available for an application to be successful, but for some applications, this may amount to nothing more than form-filling (i.e., checking which items still need to be filled in and querying the user accordingly–in any case plan inference is not yet a focal point in the current research project).

Similarly, when the system desires to "speak" to the user, it can pass a logical form representing the content to the generator. The generator then, depending on its sophistication, may access the discourse structures, to see exactly what type of utterance is preferred (for example, a good generator may use pronouns, if it can determine what discourse referents are the most salient). Again, the overall modularity of the system comes into play here, because the generator may be developed by a different set of people than those who work on the application or the parser or the dialogue system. We hope, therefore, that provided appropriate interfaces, we have created an architecture which can not only be used with a number of different task-oriented applications, but can also be continually improved upon, as each module is developed.

6.2 Summary and Future Work

We have presented what we feel is a streamlined and modular architecture for a human/computer dialogue system, and have attempted to demonstrate how the discourse structures of such a system can be used to facilitate efficient resolution of a number of phenomena which we take to be presuppositional, such as definite reference

[43] Another step that must take place in the implementation, which like many other implementation-specific details is not covered at length in this paper, is what we call *domain specialization*, from the language produced by the parser, to that used by the dialogue system and application. Domain specialization is a kind of logical coercion, which involves the specialization of general predicates into domain specific ones. For example, as mentioned earlier, application specific predicates such as *have-vacancies* won't be directly produced in the logical forms outputted by the parser, but are used in the knowledge base. This specialization may be implemented in the operators, but this sacrifices modularity (since we would not, for example, in every application always want an operator to insert a date argument any time it saw questions regarding vacancies). A better approach is to allow the application to provide a set of domain specialization rules, which the parser may access, to produce logical forms using the appropriate predicates.

and anaphora. We have also briefly described how this approach can be extended to other presuppositional phenomena, such as domain restriction and ellipsis.

In comparing our approach to SDRT, we have pointed out a number of deliberate design features in our approach (e.g., the lack of a persistent intentional model of the user) which we believe make the system more computationally tractable, and certainly more easily implemented. Indeed, these sorts of resource-saving sacrifices are often necessary in practical implementations.

Our architecture allows different developers to focus on different parts of the process. The current research aim has been to develop the overall interfaces, and to write the resolution algorithms for the individual operators. In the future, we hope to integrate our work with other modules, such as a generator, and to develop a user interface, so that the entire system may be empirically evaluated. Once this step is accomplished, we can then compare our system performance with real-world data, in an attempt to both better inform our theory of discourse and improve our system performance.

REFERENCES

ASHER, NICHOLAS. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

——, & ALEX LASCARIDES. 1998a. Questions in dialogue. *Linguistics and Philosophy* 23.237–309.

——, & ——. 1998b. The semantics and pragmatics of presupposition. *Journal of Semantics* 15.239–300.

BOHLIN, PETER, & STAFFAN LARSSON. 1999. Godis and the dialogue move engine toolkit. In *Demonstration Abstracts from the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, Maryland. Association for Computational Linguistics.

CARLSON, LAURI. 1983. *Dialogue Games: An Approach to Discourse Analysis*. Dordrecht: Reidel.

GAMUT, L.T.F. 1991. *Logic, Language, and Meaning, volume II*. Chicago, IL: The University of Chicago Press.

GINZBURG, JONATHAN. 1996. Interrogatives: Questions, facts and dialogue. In *Handbook of Contemporary Semantic Theory*, ed. by S. Lappin. Oxford: Blackwell Publishers.

GROENENDIJK, JEROEN, & MARTIN STOKHOF. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. University of Amsterdam dissertation.

GROSZ, BARBARA, & CANDACE SIDNER. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12.175–204.

GROSZ, BARBARA J., ARAVIND K. JOSHI, & SCOTT WEINSTEIN. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21.203–225.

GROSZ, B.J. 1981. Focusing and description in natural language dialogues. In *Elements of Discourse Understanding*, ed. by B. Webber A. Joshi & I. Sag, 84–105. New York: Cambridge University Press.

HEIM, IRENE. 1982. *On the Semantics of Definite and Indefinite Noun Phrases*. University of Massachusetts at Amherst dissertation.

KAMP, HANS, & UWE REYLE. 1993. *From Discourse to Logic*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

KASPER, ROBERT T., PAUL C. DAVIS, & CRAIGE ROBERTS. 1999. An integrated approach to reference and presupposition resolution. In *Proceedings of the ACL'99 Workshop on the Relationship Between Discourse/Dialogue Structure and Reference*, College Park, Maryland.

LASCARIDES, ALEX, & NICHOLAS ASHER. 1993. Temporal interpretations, discourse relations, and commonsense entailment. *Linguistics and Philosophy* 16.437–493.

MACGREGOR, ROBERT M. 1991. Using a description classifier to enhance deductive inference. In *Proceedings of the Seventh IEEE Conference on AI Applications*, 141–147.

MANN, WILLIAM C., & SANDRA A. THOMPSON. 1987. Rhetorical structure theory: A theory of text organization. *ISI Reprint Series* ISI/RS-87-190.

ROBERTS, CRAIGE. 1995. Domain restriction in dynamic semantics. In *Quantification in Natural Languages*, ed. by A. Kratzer E. Bach, E. Jelinek & B.H. Partee. Kluwer Academic Publishers.

——. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. In *OSU Working Papers in Linguistics, Vol 49: Papers in Semantics*, ed. by Jae-Hak Yoon & Andreas Kathol. (also available via the Internet at ftp://ling.ohio-state.edu/pub/roberts/infostructure.ps).

——. 1998. The place of centering in a general theory of anaphora resolution. In *Centering Theory in Discourse*, ed. by A.K. Joshi M.A. Walker & E.F. Prince. Oxford: Clarendon Press.

——. 2000 (to appear). Uniqueness in definite noun phrases. *Linguistics and Philosophy* . (also available via the Internet at ftp://ling.ohio-state.edu/pub/roberts/uniqueness.ps).

RUSSELL, BERTRAND. 1905. On denoting. *Mind* 66.479–493.

STALNAKER, ROBERT. 1979. Assertion. In *Syntax and Semantics, Volume 9*, ed. by Peter Cole.

SURI, LINDA Z., & KATHLEEN F. MCCOY. 1994. Raft/rapr and centering: A comparison and discussion of problems related to processing complex sentences. *Computational Linguistics* 20.301–317.

VAN DER SANDT, ROB. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics* 9.333–377.

——. 1999. Domain restriction. In *Focus: Linguistic, Cognitive, and Computational Perspectives*, ed. by Peter Bosch & Rob van der Sandt. Cambridge, UK: Cambridge University Press.

ZEEVAT, HENK. 1989. A compositional approach to discourse representation theory. *Linguistics and Philosophy* 12.95–131.

ACOUSTIC-PERCEPTUAL CORRELATES OF SENTENCE PROMINENCE IN ITALIAN*

Mariapaola D'Imperio

Abstract

Research on the acoustic correlates of perceived accentual prominence has generally focused on fundamental frequency (F0) alone, while few studies have attempted to shed light on how other parameters, such as duration and intensity, might interact with F0. A previous study on Italian lexical stress perception shows that duration has a major role. The present work reports on results of an experiment using synthetic speech to test which aspects of the signal, among F0, duration and intensity, are more influential in the perception of prominence structure at the sentence level and whether there are differences between questions and statements. To this end, a series of hybrid LPC-resynthesized stimuli were presented to 22 Italian listeners for forced-choice judgments. The results suggest a bigger impact of the hybridization on interrogative utterances.

1. Introduction

As defined here, prominence is the subjective salience of an element in an utterance. Most recent research on the acoustic correlates of perceived prominence in

*I would like to thank Nick Cipollone for his much needed help in writing the hybridization program used in this study (re-elaborated from an earlier version written by Mary Beckman). Thanks also to Keith Johnson and Jen Muller for comments on a previous version of this paper. Finally, I would like to thank Jose Benki for help with the statistics and for discussing the implications of the results.

speech has focused on F0 or pitch (e.g. Liberman & Pierrehumbert, 1984; 't Hart *et al.*, 1990; Ladd *et al.*, 1994). Comparatively few studies have attempted to shed light on the complex nature of prominence as a result of the interplay of parameters other than F0, e.g. duration and amplitude[1].

From the literature on the topic, it appears that prominence is primarily cued by the presence of a noticeable pitch change or by extreme (either high or low) pitch levels relative to the context (Pierrehumbert, 1980; Pierrehumbert and Beckman, 1988). However, it has been noted that even though pitch variations are not as marked in spontaneous speech as for read speech, clearly perceptible prominences can still be detected, which could be attributed to other physical indices, such as duration and/or amplitude (Boves, ten Have and Vieregge, 1984). More recently, Campbell (1995) has shown that in dialogue speech spectral information can compensate for the lack of tonal cues, when detecting prominence. It also appears that interesting differences exist in the perception of prominence between listeners with different linguistic backgrounds. For instance, Lehiste and Fox (1993) found a stronger effect of duration on Swedish listeners, as opposed to English listeners, in prominence perception.

The present study aims at uncovering the perceptual role of certain acoustic correlates of prominence in Italian, namely duration, amplitude and fundamental frequency. The relative salience of the aforementioned correlates has been already tested for isolated words in this language. Previous experiments (Bertinetto, 1980) aimed at discovering the relative weight of each of those correlates in determining lexical stress pattern, in minimal pairs such as *áncora* "anchor" and *ancóra* "again", but did not study prominence at the sentence level.

Another important difference with previous studies pertains to methodological issues. We are still far from understanding the complex proportional variations due to variables such as position in the utterance or natural occurring combinations of different parameters for such free manipulations to be useful. Hence, the present study attempts to overcome past methodological problems and to examine sentence level phenomena. The stimulus set employed in this work was generated through a technique that is very different from the one used in earlier experiments on prominence perception in Italian. Specifically, the correlates of prominence will not be directly manipulated here. An experiment was then designed in order to assess the weight of each of the acoustic correlates of stress in Italian, by cross-combining the acoustic substance of natural utterances where the focus, broad or narrow, is placed on different elements.

Despite the methodological discrepancies, previous research suggests that Italian subjects are very sensitive to durational differences, both in perception of lexical stress patterns (Bertinetto, 1980) as well as in the perception of unstressed syllable duration (Bertinetto and Fowler, 1989). It is plausible, therefore, that sentential prominence in Italian is cued by duration and intensity, as well as F0. We expect, then, that replacing only one prominence correlate (i.e., duration, or intensity, or F0) of a "donor" utterance with that of a "recipient" utterance will affect perceived prominence. This kind of manipulation was carried out for this study, whose details will be presented below. The results presented here suggest in fact that the role of duration is particularly important in the perception of specific intonation patterns.

[1]In this paper, I shall use the term "amplitude" and "intensity" interchangeably to refer to the physical property of the signal producing the subjective sensation of loudness.

2. Previous studies

The investigation of perceptual cues of stress goes as far back as the 1950s, when the classic experiments described in Fry (1955, 1958) were performed. Those studies found that acoustic prominence is concerned with certain physical correlates of the salient syllable in a word. This aspect of prominence is believed to be associated primarily with high degree of pitch variation, long duration and high amplitude (Fry, 1955,1958; Lieberman, 1960; Lehiste, 1970).

While Fry's studies had determined that F0 was indeed the most important correlate for stress in English, three decades later Beckman (1986) reestablished the role of intensity through the use of a loudness measure[2]. In her perception experiments with Japanese and English, she found in fact that F0 has a much greater role in Japanese than in English for the purpose of signaling stress. English listeners seemed to pay more attention to loudness differences[3].

Recently, most experiments on prominence perception have concentrated on the role of fundamental frequency (Ladd *et al.*, 1994; Terken, 1992; Hermes and Rump, 1994; Bartels and Kingston, 1996). Terken (1992) investigated the relative importance of fundamental frequency change and fundamental frequency maximum in determining prominence judgments in subsequent peaks, finding that the relation is more complex than expected. Hermes and Rump (1994), despite admitting that "the physical attribute underlying prominence perception is multidimensional" (p. 90), investigate perceptual prominence of falling and rising pitch movements while regarding intensity and duration as secondary cues that can only "intensify" an already existing accent. The authors used a method in which subjects had to adjust the pitch of an accented syllable in order to match the prominence of a previously heard accent. As was noticed by the authors, however, since the only adjustable dimension was pitch, it may well be that subjects tended to pay attention only to this cue and not to others.

In Italian, unlike English and Swedish, few perceptual experiments focusing on prominence, especially at the sentence level, have been performed. The only study that has explored the perceptual interaction of the various acoustic correlates in Italian is Bertinetto (1980). This study investigated the relative weight of duration, fundamental frequency and intensity on the perception of stress in the bisyllable [papa]. This segmental sequence can have two different meanings according to the stress pattern, i.e. "Pope" ['papa] or "daddy" [pa'pa]. Bertinetto (1980) argued that the role of duration is markedly greater than that of intensity and F0 for signaling word stress in Italian. F0 was instead found to be the weakest cue. He also found a listener bias in favoring the second syllable of the bisyllable when judging stress. This could have been a result of the

[2]This measure of loudness is actually labeled "total amplitude" in Beckman (1986) and is a measure that combines duration and amplitude.

[3]Beckman, who finds a pattern very similar to Nakatani and Aston (1978), offers an explanation for the difference of her results with Fry's findings. In Fry (1958), F0 overruled amplitude and duration as a correlate of word stress in a dramatic way. Beckman notices that the kind of synthesis used by Fry might have unnaturally reproduced intensity by simply attributing level values to the segments, without preserving naturally occurring contours and thus sounding very unnatural. Conversely, the LPC resynthesized stimuli that Beckman and the present study employ might make for more naturally sounding stimuli and, therefore, for a higher effectiveness of the amplitude parameter.

positional characteristics of the two syllables[4]. Though the results are very interesting, this study had some methodological limitations, which prevent a conclusive interpretation. Those limitations are mainly related to the issue of directly manipulating prosodic cues, which was avoided in the present study.

An additional variable introduced in this study pertains to the influence of modality in prominence perception, in other words whether questions are different from statements in this respect. Ultimately, I would like to discover whether the pitch values alone produce an overriding pattern of prominence responses or if the duration/amplitude values can, as predicted by Bertinetto's results, significantly determine the identification of the prominence pattern. Since we are not at a stage in which we can give an account of the prosodic organization of Italian, it was necessary to validate prominence patterns identified according to standard linguistic theories and to acknowledge observed patterns that do not strictly follow established theoretical beliefs.

For this purpose, a preliminary study (D'Imperio, 1997a) was designed in order to assess the perceptual prominence response of Italian subjects to natural speech stimuli varying in focus placement (early, medial, late) and focus type (broad vs. narrow). This preliminary experiment serves as background to the experiment described here, in which synthetic stimuli were manipulated. The experiment validated the robust recognition of intended focus in narrow focus utterances, while yielding results around chance for broad focus statements (while late focus was always identified as such in broad focus questions). Broad focus seems to be signaled by an accent that is less salient than the narrow focus accent, in that it is downstepped. Also, the lexical item that is associated to it is generally not chosen as the "most prominent" within the utterance (D'Imperio, 1997a). Therefore, we expect that the "weaker" perceptual prominence of broad focus accents will be enhanced when one of the acoustic cues of narrow focus utterances is combined with it. Additionally, narrow focus identification will be less robust when one of the correlates of broad focus is combined with a narrow focus utterance.

The analysis of the intonation contours presented here was carried out within the ToBI framework (Beckman and Ayers, 1994). The melody is basically decomposed into "target levels" (highs and lows), which can be thought of as the "notes" associated to some specific segmental locations.

3. Methods

3.1 Stimuli

A set of stimuli was created by using the hybrid resynthesis technique first developed by Nakatani and Aston (1978) and subsequently adopted by Beckman (1986) and Hirschberg and Ward (1993). The technique consists in, first, sampling RMS amplitude, timing, LPC coefficients and pitch information for each original utterance of each stimulus pair and then synthesizing new files in which one of the sampled features of the original utterances was exchanged for those of another (with synthetic files produced by linearly interpolating between sample points). As a last step, new utterances are resynthesized on the basis of the "hybrid" files using LPC resynthesis.

[4]As it turns out, final stressed syllables appear to be shorter than syllables in other positions in production studies.

Direct manipulation of the stimuli was, as mentioned above, avoided, since it is impossible at this point to estimate parameter intervals that would be equal as to perceptual effect. The hybridization technique allows one to avoid the risk of involuntarily creating discrepancies in step sizes that would make the perception effect of one acoustic dimension seem stronger than it is in reality.

Stimuli consisted of simple Subject-Verb utterances, using the sentence *Mario esce* "Mario goes out". The original utterances were identical from a segmental point of view, while various intonational combinations of modality (questions or statements) and focus type were superimposed on them. The utterances were all produced by a female speaker of the variety of Italian spoken in Naples (the author).

As shown in Table 1, the same sentence was uttered as either a neutral utterance with broad focus (Broad) or as a narrow focused utterance, where the focus occurred on either the Subject (NarrowS) or the Verb (NarrowV). The utterances were all auditorily transcribed to check for intended focus pattern. 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+.

Mario esce "Mario goes out"	broad focus (Broad)
MARIO esce "MARIO goes out"	narrow focus on S (NarrowS)
Mario ESCE "Mario GOES OUT"	narrow focus on V (NarrowV)

Tab. 1 Patterns of sentence stress in the test utterances.

For the hybrid resynthesis, spectral coefficients of the natural utterances were obtained through an 18th-order LPC (Linear Predictive Coding), while amplitude and fundamental frequency values were extracted using an autocorrelation F0-tracking program. The values obtained were used to create hybrid utterances where just one of the acoustic correlates of prominence was exchanged at a time. For instance, the F0 donor utterance could be *MARIO esce?*, with nuclear (i.e., the most prominent accent in the sentence) accent on the subject (see Figure 2, middle), while the duration and (RMS) amplitude donor utterance would be *Mario ESCE?* (see Figure 2, lower), with nuclear accent on the verb. In such a case, the goal is to find which word will be judged the most prominent by the listeners, i.e. whether F0 cues or duration and intensity cues will have a stronger impact in this sense.

Non-hybrid	**F01**+LPC1+RMS1+D1
F0 change	**F02**+LPC1+RMS1+D1
RMS change	**RMS2**+F01+LPC1+D1
Dur. Change	**D2**+F01 +LPC1+RMS1

Tab. 2 Acoustic correlate manipulations used in the Experiment. 1 = donor utterance; 2 = base utterance.

The order of the base utterance/donor utterance combination could be reversed to allow for indirect exchange of original spectral parameters. For example, the fundamental

frequency of a broad focus utterance was in one case combined with spectral, amplitude and duration values of values of a narrow focus utterance (either on the subject or on the verb). In another instance, the fundamental frequency of the narrow focus utterance was combined with spectral, amplitude and duration values of the broad focus utterance. Along the same lines, the amplitude or duration of the donor utterance was in another instance combined with all other acoustic values of the base utterance. For example, as a result of inserting the fundamental frequency of the broad focus *Mario esce* in the narrow focus *MARIO esce*, with focus on *Mario*, we obtain that the stressed syllable *Ma-* (of *Mario*) will be strongly marked by the substantive values of duration and amplitude, but will not be marked by a strong pitch accent. All combinations of broad focus utterance plus one of the features of narrow focus utterances (and viceversa) were obtained. Narrow focus utterances were never combined with each other, since this produced unnatural effects.

Syllabic boundaries were marked in the original utterances, yielding 4 cuts or "anchors" (one for each syllable). Frame numbers were obtained for each cut. When duration was the parameter exchanged, the frame number for each cut in the base utterance was exchanged for the frame numbers of the donor utterance, while a linear interpolation algorithm was used to obtain new spectral, amplitude and F0 values in the hybrid utterance. When amplitude or fundamental frequency values were taken from the donor utterance, those were interpolated to the frame number relative to the anchors in the base utterance.

The original spectral coefficients were recombined with adjusted amplitude and F0 contours or simply readjusted as to frame number. Hybrid utterances were then resynthesized through LPC resynthesis. The spectral coefficients of the hybrid utterance were always derived from the base utterance and the only permissible combination was broad focus plus narrow focus utterance, and neither broad-broad nor narrow-narrow combinations were employed.

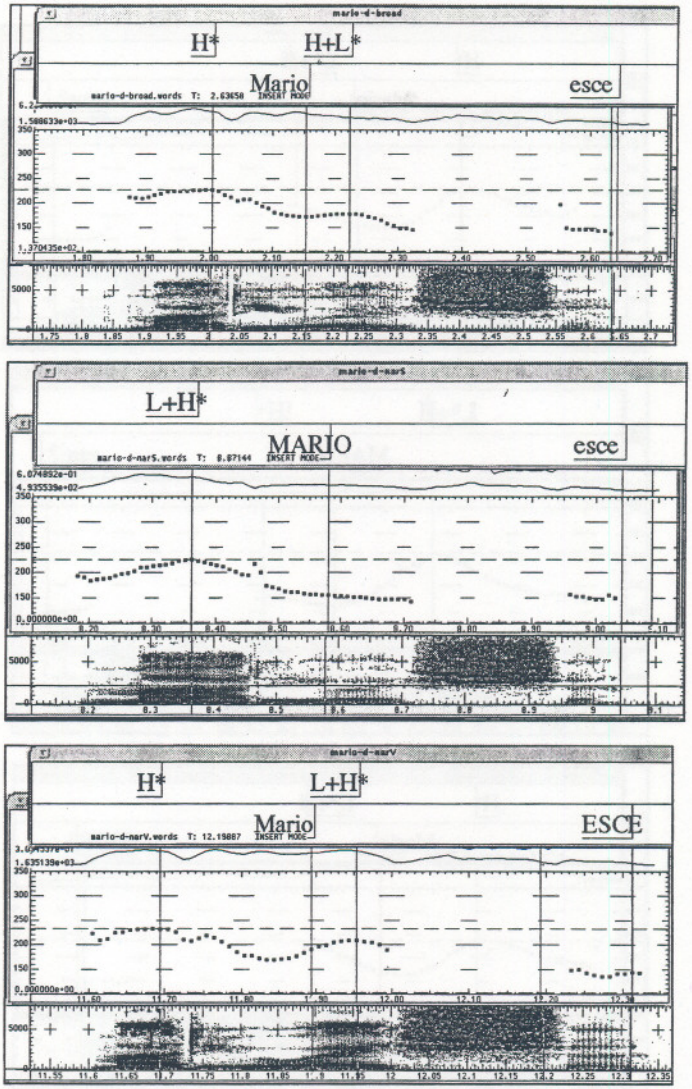


Figure 1. F0 curves and spectrograms for a broad focus declarative (upper), a declarative with narrow focus on the subject (middle) and a declarative with narrow focus on the verb (lower).

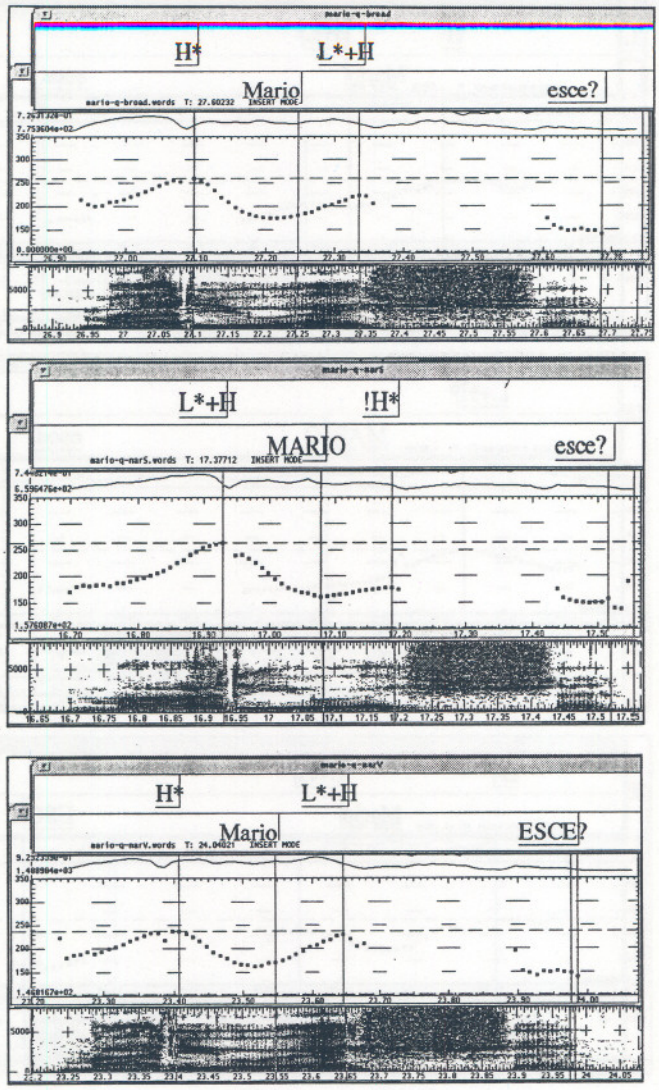


Figure 2. F0 curves and spectrograms for a broad focus question (upper), a question with narrow focus on the subject (middle) and a question with narrow focus on the verb (lower).

3.2 Procedure

The 24 hybrid stimuli plus 6 non-hybrid resynthesized originals, used as controls, were presented to the listeners in random order. The task consisted of choosing the "most important" word in the utterance (forced choice judgment) by clicking on its orthographic version presented on a computer screen. After the choice was made, the computer played the subsequent stimulus, and the following choice was made.

The listeners were instructed to listen carefully to each sentence and to choose the answer as quickly as possible after listening to each stimulus, even when not entirely sure about it. Explicit use of proper linguistic terms such as "prominence" and "focus" was avoided in order 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 Tab. 1) and had the subject point at one of the words as being the most important. The experiment was self-paced, and each stimulus was played only after the previous choice was made.

3.3 Listeners

Twenty-two listeners participated in the experiment. All but two of the listeners were undergraduate students at the University Federico II of Naples, with ages varying between 22 and 27. They were all speakers of Neapolitan Italian and hence had the same geolinguistic background of the speaker who produced the stimuli[5]. They all had normal hearing and performed the task without problems. Some of the subjects had attended introductory linguistic courses.

4. Results

The listening test yielded a total of 3300 responses (30 stimuli * 5 repetitions * 22 subjects). Three factors were used in the repeated measure Analysis of Variance (ANOVA), i.e. MODALITY, MANIPULATION and FOCUS TYPE (see Table 3). MODALITY had two levels (question vs. statement intonation), while FOCUS TYPE had 4 levels. These levels are the result of dividing up the hybrid stimuli as follows: broad focus utterance plus one of the correlates of utterances with narrow focus on V (Broad+NarrowV), broad focus utterance plus one of the correlates of utterances with narrow focus on S (Broad+NarrowS), utterance with narrow focus on S plus one of the correlates of utterances with broad focus (NarrowS+Broad) and utterance with narrow focus on V plus one of the correlates of utterances with broad focus (NarrowV+Broad). The natural utterances were grouped with the hybrid ones, according to focus type. MANIPULATION had four levels, according to the parameter that was manipulated (Duration, F0, amplitude, non-hybrid). Therefore, the design was a 2x4x4 factorial. The variables were manipulated within subjects. The number of judgments favoring verb prominence for each stimulus was determined, henceforth NUMBER OF V JUDGMENTS, which was the dependent measure. Planned comparisons were also carried out on relevant scores.

[5]Only three of the subjects were knowledgeable in linguistics, but none was aware of the purpose of the experiment.

Factors	*Levels*
MODALITY	Question, Statement
FOCUS TYPE	Broad+NarrowV, Broad+NarrowS, NarrowS+Broad, NarrowV+Broad
MANIPULATION	Duration, F0, RMS amplitude, non-hybrid

Tab. 3 Factors and levels of the statistical analysis.

In Table 4 the main effects and interactions of MODALITY, MANIPULATION and FOCUS TYPE are given.

Effects	F	P-value
Main effects		
Modality	60.37	<0.01
Focus Type	140.5	<0.01
Manipulation	8.34	<0.01
Two-way interactions		
Modality * Focus Type	0.3	NS
Modality * Manipulation	3.6	.02
Focus Type * Manipulation	50.8	<0.01
Three-way interaction		
Modality * Focus Type * Manipulation	4.49	<0.01

Tab. 4 Main effects and interactions of MODALITY, MANIPULATION and FOCUS TYPE.

The results support the hypothesis that acoustic manipulation can affect the perceived intended focus of the base utterance. A large main effect of both FOCUS TYPE and a main effect of MODALITY were found. Moreover, a significant interaction of MANIPULATION with FOCUS TYPE and a significant three-way interaction were found.

4.1 Statements

Figures 3 and 4 show the mean overall results for the four focus types. The bars in the two figures are the mean for NUMBER OF V JUDGMENTS for hybrid stimuli (duration, F0 and RMS amplitude) vs. non-hybrid stimuli. The manipulations associated with the different labels are shown in Table 5.

HYBRID STIMULUS	BASE	DONOR
Broad+S	broad focus	NarrowS
Broad+V	broad focus	NarrowV
NarrowS+B	NarrowS	broad focus
NarrowV+B	NarrowV	broad focus.

Tab. 5 Combinations of Base and donor utterances used to create the hybrid stimuli.

The results were averaged across subjects. Overall, statements present a mean score that is never greater than 4, while questions have higher values. The effect of modality was nearly significant in the two-way interaction with manipulation.

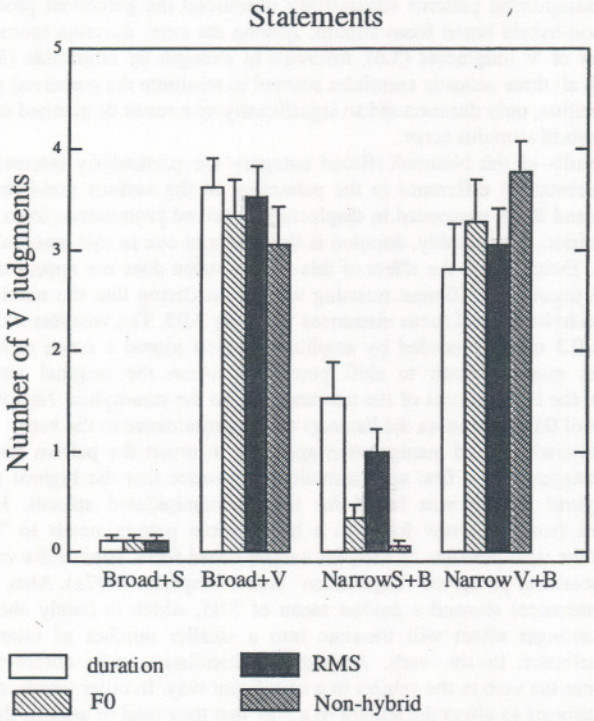


Fig. 3 Mean values for "Number of V judgments" for all speakers across focus types for statements. Manipulation levels are indicated by different bar patterns.

Figure 3 shows the mean values of the dependent variable across different focus types. The three acoustically manipulated patterns scored equally well in the Broad+NarrowS (Broad+S) manipulation, with a mean V judgement score of 0.09. This value was very close to the non-hybrid focus-S pattern, which was 0.05. In other words, all three acoustic correlates successfully displaced perceived prominence from the verb to the subject position. This might be due to the high sensitivity to the beginning of the utterance that has already been found in perception of natural utterances with varying focus position (D'Imperio, 1997a). This triplet must be contrasted to the natural broad focus stimulus in the Broad+NarrowV series. Broad focus stimuli were conventionally grouped with stimuli in which a prominence shift towards the verb was expected. In standard phonological theory, broad focus sentences have late prominence and no naturally occurring broad focus utterances have focus on S.

The Broad+NarrowV manipulation scored in the opposite direction. The three acoustically manipulated patterns successfully reinforced the perceived prominence on the verb for non-hybrid broad focus stimuli. Among the cues, duration scored a slightly greater number of V judgments (3.6), followed in strength by amplitude (3.5) and F0 (3.05). Though all three acoustic correlates seemed to reinforce the perceived prominence on the verb position, only duration did so significantly as a result of planned comparisons with the non-hybrid stimulus score..

The results of the NarrowS+Broad category are particularly interesting in that they show a substantial difference in the patterning of the various non-hybrid stimuli. Only duration and RMS succeeded in displacing perceived prominence from the subject to the verb position. Remarkably, duration is the strongest cue in this manipulation, with a mean of 1.5. Even though the effect of this manipulation does not appear unusual at a first glance, it acquires a different meaning when considering that the maximum value reached by non-hybrid broad focus utterances was only 3.05. The weakest cue appears to be F0, with a 0.3 mean, preceded by amplitude, which scored a mean equal to 1. As expected, it is more difficult to shift perception when the original utterance has prominence on the first element of the utterance. As to the non-hybrid NarrowS stimuli, only in a mean of 0.05 utterances did listeners assign prominence to the verb.

The NarrowV+Broad manipulation appeared to revert the pattern established in the previous category. At a first approximation, we notice that the highest mean score among the hybrid stimuli was found for the F0 manipulated stimuli. However, a successful shift from a narrow focus to a broad focus pattern needs to "lower" the prominence at the verb position. In fact, for natural broad focus stimuli the verb location receives low scores of perceived "importance" (see D'Imperio, 1997a). Also, non-hybrid broad focus utterances showed a modest mean of 3.05, which is barely above chance. Therefore, a stronger effect will translate into a smaller number of utterances with assigned prominence to the verb. Among the correlates, only duration displaced prominence from the verb to the subject in a significant way. In other words, the duration manipulation appears to affect the stimuli in a way that they tend to assume the uncertain prominence pattern already recorded for natural broad focus utterances (D'Imperio, 1997a). The results for stimuli with duration manipulation indeed show a mean score of 2.8, which goes in the direction of a weaker prominence on the verb. The F0 manipulation was the least different from the non-hybrid NarrowV manipulation. Amplitude (RMS) results are intermediate between the other two manipulations.

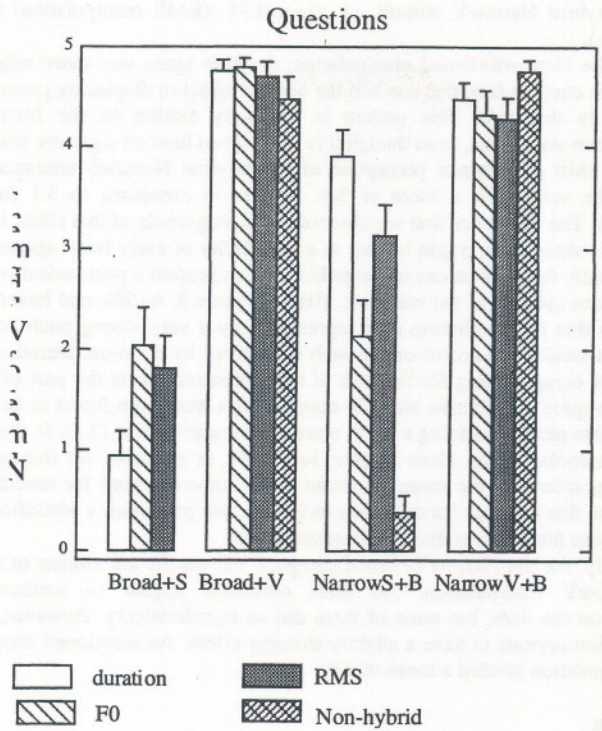


Fig. 4 Mean values for "Number of V judgments" for all speakers across focus types for questions. Manipulation levels are indicated by different bar patterns.

4.2 Questions

As Figure 4 shows, the role of duration and RMS was quite remarkable, at least in some manipulations. Within the Broad+NarrowS hybrid manipulation, duration was strongest in displacing perceived prominence from the verb to the subject position (lower bars indicate low scores of V responses and, as a result, high scores of S responses). After duration, the second highest effect is due to amplitude, followed by F0. Although it was the least effective cue, F0 scored better than chance in shifting prominence perception.

The Broad+NarrowV manipulation presents an interesting "tie" among the acoustic cues. All three hybrid levels seemed to score marginally better than the non-hybrid, broad focus stimulus, as expected, since each cue has the effect of reinforcing the perceived prominence on the verb. However, this was only a non-significant trend. Non-hybrid stimuli scored a mean of 4.5, which was lower than the 4.8 scored by non-hybrid stimuli with NarrowV focus. All hybrid stimuli registered a mean score that is very close

to the non-hybrid NarrowV stimuli, i.e. from 4.71, (RMS manipulation) to 4.81 (F0 manipulation).

For the NarrowS+Broad manipulation, duration again was more salient than the other acoustic cues. In fact, this cue had the biggest effect in displacing prominence from the subject to the verb. This pattern is strikingly similar to the NarrowS+Broad manipulation in statements, even though it is reproduced here on a greater scale. Duration was able to shift prominence perception of the original NarrowS utterances from the subject to the verb, with a score of 3.9. This score compares to 3.1 for the RMS manipulation. The difference that we observe in the magnitude of this effect in questions, as opposed to statements, might be due to a peculiarity of early focus questions. Unlike statements, early focus questions in Neapolitan Italian present a postnuclear pitch peak on the last stressed syllable of the utterance. (!H* in Figure 2, middle and lower panel). We observe here that the conflicting cues represented by a very strong pitch accent on the subject and a weak pitch accent on the verb is resolved by duration. Duration appears to be capable of compensating for the lack of tonal prominence on the part of the nuclear pitch peak in quite an effective way. F0 manipulation was again found to be the poorest correlate in this pattern, yielding a mean score that is near chance (2.1). In the non-hybrid NarrowS manipulation the mean is quite low (0.4), as expected for this pattern, even though it is greater than the mean we found in the same category for statements (0.05). This might be due to a bias for questions to receive late prominence identification due to the conspicuous pitch accent that characterizes them.

Finally, for the NarrowV+Broad category, the results are similar to those of the Broad+NarrowV manipulation. All three correlates appear to weaken perceived prominence on the verb, but none of them did so significatively. However, within this group, duration appears to have a slightly stronger effect. As mentioned above, the non-hybrid manipulation yielded a mean of 4.8.

5. Discussion

The results appear to support the hypothesis that duration is an important correlate of prominence in Italian, not only at the word level (Bertinetto, 1980), but also at the sentence level. At least for two manipulations, i.e. Broad + NarrowS donor and NarrowS + Broad donor, duration is the correlate that has the biggest impact in displacing perceived prominence. In all of these manipulations, F0 is the weakest cue, which parallels Bertinetto's findings[6].

The results provide strong support for the idea of a trading relation among acoustic cues in the perception of prominence. The hybrid Broad+NarrowS manipulation completely reverted the prominence pattern of broad focus base utterances, for instance. Such manipulation had the effect of making listeners assign prominence to the subject most of the time, in both question and statement stimuli. Moreover, when NarrowS base questions where combined with a broad focus question as a donor (NarrowS+Broad manipulation), prominence was significantly shifted to the verb (except for the F0

[6]Bertinetto's view of duration contribution has to be seen in the right perspective, though: "Thus, although D undoubtedly bears the greatest importance in the determination of perceptive responses concerning prominence, this component must not be viewed separately from the others. When certain conditions are met, the combined effects of I and F0 may in fact exceed the weight of D" (Bertinetto 1980, p. 392).

manipulation). The hybrid manipulation also had the effect of reinforcing prominence on the verb in the Broad+NarrowV manipulation and reducing it in the NarrowV+Broad manipulation. This result was true only for statements and was expected from the typical prominence responses to non-synthetic broad and late narrow focus stimuli (D'Imperio, 1997a).

The NarrowS+Broad manipulation had a bigger overall effect in interrogatives. This is probably due to the different postnuclear contour of early focus interrogatives as opposed to early focus declaratives (D'Imperio, 1997b).

The present results can be compared to Beckman's (1986) results for American monolingual subjects. When separately looking at amplitude and duration in the American-monolinguals results, duration was more effective than amplitude. However, in Beckman (1986) the most effective cue overall was F0.

One outcome of the present experiment that cannot be compared to previous studies is the effect of modality. Especially interesting is the comparison between Broad+NarrowS statements and questions. While in the statements all three manipulations produced a very strong effect, in the questions they did not. In fact, F0 and intensity did not succeed in shifting prominence perception in this condition as successfully as duration did. This outcome can be explained by the fact that, unlike statements, NarrowS questions present a late postnuclear pitch-accent (see middle and lower panel of Figure 2; see also D'Imperio, 1997b). In this case, switching the melodic contour of a NarrowS question has the effect of slightly decreasing the percept of a tonal event on the verb, which could account for the weaker effect of F0. In statements, the melodic contour of a NarrowS utterance has no postnuclear tonal markings (see Figure 1, middle and lower panel); therefore, no late tonal event can attract perceptual prominence.

The effect of duration in the NarrowS+Broad questions is even more surprising in the light of what we know about preboundary lengthening (Beckman and Edwards 1990), by which the phrase-final section of an utterance is lengthened. Just as it appears that listeners can factor out the gradual declination of F0 in the course of an utterance, it is also expected that they would perceptually adjust for longer utterance portions in the proximity of a boundary. However, this was not the case in the question results. The percept of a longer verb constituent made it perceptually more prominent than the pitch prominent subject. In this case, the duration of the stressed vowel in the first syllable of *esce* traded for the lack of a perceptually strong pitch accent for the purpose of signaling prominence on that word. The strength of the duration manipulation is further supported by the higher consistency in the results for this manipulation as opposed to the F0 and RMS manipulation (see § 4.2 above).

What these results mean for traditional trading relation hypotheses is difficult to say for a number of reasons. First, most of the literature on the topic of the last decade has concentrated on segmental features[7], like the feature [voice] or manner features such as [fricative] (see Repp 1982 for a review). Auditory integration can be evoked to explain the trading relation by appealing to psychophysical properties of the auditory system

[7]We also know that prominence (or stress) is not a feature, at least not in the sense as [+ voice] or [-velar] are. Since Liberman (1977) our view of metrical strength has changed from being an absolute, categorical, value (as in Chomsky and Halle, 1968) to a relational dimension between terminal elements in a structure. It may well then be that it is not possible to easily generalize from feature perception to prominence perception and that the two fields have to be kept apart.

(Kingston and Diehl, 1995). This position is severely criticized by motor theory supporters such as Repp (1982), who claims that (p. 93) "In most other cases, [however], the cues that participate in a trading relation are simply too diverse or too widely spread out to make auditory integration seem plausible" (brackets inserted by the author). The extreme position represented by Repp is one that simply denies cue integration as a generic auditory process and which, instead, regards it as yet another proof of the existence of a specialized phonetic mode of speech perception. In this perspective, trading relations among acoustic cues could only occur "because listeners perceive speech in terms of the underlying articulation and resolve inconsistencies in the acoustic information by perceiving the most plausible articulatory act" (Repp, 1982: p. 95).

In order to support a speech-specific view of trading relations in the realm of prosody, motor theorists can appeal to the results of works such as Smith (1978), cited in Repp (1982). In this study, relative duration of two subsequent syllables was varied and two types of judgments were elicited from the subjects, one linguistic (stress position) and the other auditory (which syllable was longer). It was found that subjects had a first syllable bias only when they were performing the linguistic task. The explanation given to account for the bias is that, when listeners are in a "speech mode" of perception, they expect the second syllable to be longer because of the speech specific phenomenon known as final lengthening. In other words, when perceiving the stimuli as speech, longer duration in the second syllable is not as strong a cue as in the first syllable, hence a first syllable bias in the responses. Bertinetto's (1980) results point to something similar, though in the opposite direction. In this work, subjects showed a second syllable instead of a first syllable bias. The suggested explanation is that they might have adjusted for the intrinsic shorter length of final stressed syllables reported in Italian production studies.

A speech-specific interaction of prosodic correlates of stress is argued against by Beckman (1986). On one hand, in her results intensity and duration could be seen as being in a trading relationship because of their common articulatory origin: an augmented jaw movement can result in a longer as well as in a more intense acoustic signal[8]. However, drawing from psychoacoustic literature on temporal summation of loudness, Beckman proposes that the special relationship between intensity and duration has an auditory and not simply an articulatory basis. Loudness, in fact, appears to be the result of the combined effect of intensity and duration over a segment (see psychoacoustic literature cited in Beckman 1986). The claim that duration contributes to the loudness percept in speech has been recently opposed by Sluijter *et al.* (1997): "It has only been established for pure tones of a relatively short duration that differences in duration are responsible for differences in the perception of loudness." (p. 511). Sluijter *et al.* argue instead for a relevant effect of intensity manipulations over high frequency regions of the spectrum. Mere intensity level (i.e. affecting the entire spectral range) variations are regarded, instead, as having no "communicative significance" because of their vulnerability to environmental masking. Intensity is, in fact, highly affected by environmental noise, position of the mouth, intervening obstacles, etc. However, the role of intensity level as expressed by RMS amplitude in the present study cannot be entirely

[8]Another complication of duration as a cue to stress is due to its ambiguous articulatory origin. In Articulatory Phonology terms, longer duration can be a result of either reduced stiffness in the gesture or a result of changes in intragestural phasing (Browman and Goldstein, 1990). Our data cannot say anything about this matter, since it is impossible to differentiate between the two hypotheses on

dismissed. Amplitude manipulations appeared, in fact, to have an effect that was even stronger than the F0 manipulation in the majority of cases.

It is interesting that in the present experiment conflicting cues did not give rise to extremely confused results, as direct realism theory would predict (see Fowler, 1996 discussion of the results of Fitch *et al.*, 1980). Since no discrimination task followed the forced identification, no strong argument against this view can be provided at this point. However, the findings presented here appear to speak against a direct realist view of speech perception for additional reasons. If prominence is directly perceived, we would have to postulate a unique articulatory gesture decoded from the acoustic proximal event. The problem is that while it is somehow possible to postulate a common origin of intensity and duration variations, it is more difficult to reconcile the articulatory production of these last two cues with fundamental frequency production. In other words, both increased laryngeal activity and jaw opening, say, should be both translated back to the linguistic category "prominent". Should we favor a more abstract motor theoretic approach, we could hypothesize that what listeners do is decode some kind of speech "effort" localized on the prominent syllables (see de Jong (1995) for articulatory characteristics of stressed syllables). This "effort" can be translated back to either neural commands for jaw opening, subglottal pressure increase or greater laryngeal activity, or, alternatively, to a combination of them.

It seems to me that the best explanation for the data presented here is the "strong auditorist" perspective represented by works such as Kingston and Diehl (1995). This view entails that some acoustic properties cohere not just when sharing a common articulatory origin, but also when producing the same auditory effect. In other words "certain acoustic correlates of a phonological distinction are integrated into perceptual properties that enhance contrasts" (Kingston and Diehl, 1995, p. 24). It may also be that cues are integrated into an *intermediate perceptual property* (IPP), which in this study would be the percept of something being *prosodically stronger*. That the cues enhance each other is proven by the results of the broad focus + late (narrow) focus manipulations. In order to prove the soundness of the theory, we would need to perform a test where synthetic stimuli, sufficiently different from speech, would be used. Moreover, we would still have to account for the language-specific nature of the postulated IPP level. An alternative proposal, as suggested in Nearey's commentary on Kingston and Diehl's paper (Nearey, 1995) is that the IPPs are actually relevant only in the process of language acquisition and that we need not postulate them as independent levels in the representation. The problem is that our knowledge of psychoacoustic cue integration cannot be easily applied to language (for instance, one cannot easily extend the findings on pure tone perception; cf. Sluijter *et al.* 1997 criticism presented above).

6. Conclusion

The hybridization method appears to successfully affect perceived prominence in Italian. Specifically, duration appears to have a dominant role when the "donor" and "recipient" utterance have different accent structure (as in Broad+NarrowS manipulations. Differences in overall accent structure between questions and statements seem to determine differences in the effect of the manipulation. Our results present a

problem for theories where pitch is the primary correlate of prominence[9]. The results support a view by which duration is an active prominence cue in nuclear stress perception in Italian, and, more broadly, represent a crucial step towards understanding the interplay of language-specific acoustic correlates of stress.

REFERENCES

Bartels, C. and Kingston, J. (1995). *Salient pitch cues in the perception of contrastive focus*, in Dickey, M.W. and S. Tunstall (Eds.) UMOP 19, pp. 1-25.

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

Beckman, M.E. and Ayers, G. M. (1994). *Guidelines for ToBI Labelling*. Unpublished manuscript, Ohio State University. [Send email 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 Edwards, J. (1990). *Lengthenings and shortenings and the nature of prosodic constituency*, in J. Kingston and M.E. Beckman (eds.), *Papers in Laboratory Phonology II*, Cambridge, CUP, pp. 152-178.

Bertinetto, P.M. (1980). *The perception of stress by Italian speakers*, J. of Phon., 8, pp. 385-95.

Bertinetto, P.M. and Fowler, C.A. (1989). *On sensitivity to durational modifications in Italian and English*, Rivista di Linguistica, 1, 1, pp. 69-94.

Boves, L., Ten Have, B.L., Vieregge W.H. (1984). *Transcription of Intonation in Dutch*, in Gibbon., D. and H. Richter (eds.), *Intonation, Accent and Rhythm*, Berlin, De Gruyter, pp. 20-45.

Browman, C.P. and Goldstein, L. (1990). *Gestural specification using dynamically-defined articulatory structures*, J. of Phon., 18, pp. 299-320.

Campbell, W.N. (1995). *Loudness, spectral tilt and perceived prominence in dialogues*. Proc. ICPhS 95, vol. 3, pp. 676-679.

D'Imperio, M. (1997a). *Breadth of focus modality and prominence perception in Italian.* OSU Working Papers in Linguistics, 50, pp. 19-39.

D'Imperio, M. (1997b). *Narrow focus and focal accent in the Neapolitan variety of Italian.* Proceedings of an ESCA Workshop on Intonation, Athens, pp. 87-90.

D'Imperio, M. (1998). *Prominenza accentuale, focus e modalità intonativa nella percezione di parlato italiano letto*. In Proceedings of the "VIIIe Giornate di Studio del Gruppo di Fonetica Sperimentale (GFS)", December 18-20, Pisa, Italy.

de Jong, K.J. (1995). *The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation*, JASA, 97, pp. 491-504.

[9]Beckman (1996), p. 38 "For example, accented vowels tend to be longer and articulated closer to the periphery of the vowel space (see de Jong, 1995, for a review and some recent data). However, these are minor variations compared with the qualitative difference between inherently longer full vowels and inherently very short reduced vowels that categorically defines the stress contrast between heavy and light syllables at the lowest level of the stress hierarchy, and could be called ancillary to the tonal markers (Beckman & Edwards, 1994). Thus it is not possible to talk about stress at the two higher levels without explicitly or implicitly assuming an intonational pattern for an actual or imagined utterance of the text" (the boldface is mine).

Farnetani, E. and Kori, S. (1983). *Acoustic manifestation of focus in Italian*, "Quaderni del Centro di Studio per le Ricerche di Fonetica", 2:287-318.

Fowler, C. (1996). *Listeners do hear sounds, not tongues*, JASA, 99 (3), pp. 1730 41.

Fry, D.B. (1955). *Duration and intensity as physical correlates of linguistic stress*. JASA, 23, pp. 765-769.

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

't Hart, J., Collier, R. and Cohen, A. (1990). *A perceptual study of intonation*. Cambridge, England, CUP.

Hermes, D.J. and Rump, H.H. (1994). *Perception of prominence in speech intonation induced by rising and falling pitch movements*. JASA, 96 (1), pp. 83-92.

Hirschberg, J. and Ward, G. (1992). *The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English*, J. of Phon. 20, pp. 241-51.

Kingston, J. and Diehl, R.L. (1995). *Intermediate properties in the perception of distinctive feature values*, in Connell, B. and A. Arvaniti (eds.) *Papers in Laboratory Phonology IV*, Cambridge, CUP, pp. 7-27.

Ladd, R.D., Verhoeven, J. and Jacobs, K. (1994). *Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects*, J. of Phonetics, 22:87-99.

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

Lehiste, I. and Fox, R.A. (1993). *Influence of duration and amplitude on the perception of prominence by Swedish listeners*, Speech Communication 13, pp. 149-54.

Liberman, M. and Pierrehumbert, J.B. (1984). *Intonational invariance under changes in pitch range and length*, in M. Aronoff & R.T. Oehrle (eds.), *Language Sound Structure: Studies in phonology*, 157-233, Cambridge, MA: MIT Press.

Lieberman, P. (1960). *Some acoustic correlates of word stress in American English*, JASA, 22, pp. 451-454.

Nakatani, L. and Aston C. (1978). *Acoustic and linguistic factors in stress perception*. Unpublished manuscript, Bell Laboratories.

Nearey, T. M. (1995). *A double-weak view of trading relations*, in Connell, B. and A. Arvaniti (eds.) *Papers in Laboratory Phonology IV*, Cambridge, CUP, pp. 28-40.

Pierrehumbert, J.B. (1980). *The phonology and phonetics of English intonation*. Doctoral dissertation, MIT, Indiana University Club.

Pierrehumbert, J.B. and Beckman, M.E. (1988). *Japanese Tone Structure*. Cambridge, MA, MIT Press.

Repp, B.H. (1982). *Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception*, Psychological Bulletin, vol. 92 (1), pp. 81-110.

Sluijter, A.M.C., van Heuven, V.J. and Pacilly, J.J.A. (1997). *Spectral balance as a cue in the perception of linguistic stress*, JASA, 101 (1), pp. 503-13.

Terken, J. (1992). *Fundamental frequency and perceived prominence of accented syllables*. JASA, 89 (4), pp. 1768-76.

AN AUTOSEGMENTAL/METRICAL ANALYSIS OF SERBO-CROATIAN INTONATION *

Svetlana Godjevac

Abstract

Based on the qualitative analysis of the F_0 contours of wide range of utterances (broad focus declaratives, broad focus questions, narrow focus declaratives, narrow focus questions, vocative chant, and prompting intonation) utterred by nine native speakers, an autosegmental/metrical analysis of Standard Serbo-Croatian intonation is proposed. This analysis argues for sparse specification of tones, contra Inkelas and Zec (1988), and two levels of prosodic phrasing: the phonological word and the intonational phrase. The phonological word is defined in terms of a lexical pitch accent and an initial word boundary tone, whereas the intonational phrase is a domain defined by pitch range manipulations (expansion, compression, reset, downstep) and final intonational phrase boundary tones.

1 Introduction

Standard Serbo-Croatian (SC) is a pitch-accent language. All analyses (Browne & McCawley 1965 (B&M), Inkelas & Zec 1988 (I&Z), Kostić 1983, Lehiste & Ivić 1963,

*I would like to express my gratitude to Mary Beckman, Chris Barker, Allison Blodgett, Rebecca Herman, Molly Homer, Tsan Huang, Ilse Lehiste, Gina Taranto, and Pauline Welby. I also wish to thank my informants: Dragana Aleksić, Ljubomir Bjelica, Ana Dević, Ksenija Djuranović, Svetislav Jovanović, Jasna Kragalott, Svetlana Likić, and Branislav Unković for their patience and kindness in providing the data. All errors are mine.

1986 (L&I), Nikolić 1970, Stevanović 1989, Gvozdanović (1980), inter alia) recognize four different types of accents: short falling, long falling, short rising, and long rising. In this paper I present an analysis of surface tones of these accent types in different sentential environments, including broad-focus and narrow-focus utterances, citation form, vocative chant, prompting intonation, and questions. This analysis is based on the instrumental study of recorded utterances by eight native speakers. It is an autosegmental/metrical analysis because the F_0 shapes are decomposed into their component parts, and the tones and the backdrop pitch range are analyzed in terms of their relations to metrical structure.

The general observation differentiating this proposal from earlier autosegmental accounts, is that even the surface tones in SC are sparsely specified to moras, the tone bearing unit in SC. More specifically, the analysis argues for three main innovations over the cited analyses: (i) a decomposition of word tone strings into a demarcative tone, a boundary tone, and accent proper (rather than H-tone spreading and default L-insertion); (ii) bitonal accents with the initial tone starred (i.e. associated to the accented syllable) and the trailing tone unassociated; and (iii) no neutralization of the lexical accents in declarative sentence final position. My proposal regarding SC prosodic structure includes two prosodic units: a phonological word and an intonational phrase. Their tonal properties are defined in terms of specification of accents, boundary tones, and pitch-range manipulation. In addition, some observations of more global pitch trends, such as downstep, are offered.

One reason a refined picture of SC word tones is important is that it serves as the foundation of an ongoing study of the interaction of intonational effects such as pitch range compression and downstep with syntactic scrambling, word-order focus, etc. These in turn are central to interpretation. The interaction of intonation with interpretation is left for future study. More immediately, this study serves to add to descriptions of prosodic structure of pitch accent languages, which include Japanese, Norwegian, and Swedish, thereby contributing to the crosslinguistic study of variation in prosody.

I argue that SC's four accents are bitonal. The falling accents are H*+L, whereas the rising accents are L*+H, where '*' marks the tone associated with the relevant tone bearing unit within the stressed syllable, as in Bruce's (1977, 1990) analysis of Swedish word accent. The consequence of this proposal is that the second tone is not linked to a particular mora but is phonologically unassociated. As we will see, a long falling accent may realize both tones on the stressed syllable, whereas in words with a short falling accent, the trailing tone is usually realized on the poststressed syllable, and sometimes is even truncated.

Not all words in SC carry a pitch-accent. Verbal and pronominal clitics, prepositions, and most conjunctions do not bear pitch-accents.[1] These words cliticize to an adjacent word which does bear a pitch-accent to form a phonological word. A phonological

[1]Zec & Inkelas (1990) assume that the division between phonological words and clitics aligns with the syntactic division into content and function words. This division seems generally right but there are a few

word is the smallest prosodic unit, and is tonally marked by a pitch accent. (As we'll see, proclitics are marked by a L word boundary tone (which I will mark as %L), but they lack a pitch accent.) In the case of SC, I propose, the relevant tonal marking is a pitch-accent and a %L word boundary tone. As a general rule, there is maximally one pitch accent and one %L word boundary tone per phonological word. (As will be discussed in section 4.1.2, there are exceptions to this rule. Some polymorphemic words can be realized with two pitch-accents, but they are in free variation with variants realized with one pitch accent. Proclitics also bring an additional word boundary tone.) That is, a phonological word in SC has exactly one head syllable (marked with lexical pitch-accent) and at least one edge (word boundary) marked tonally.

The sentential tune in a declarative utterance under broad focus shows an overall downtrend in the pitch level. (By broad focus I mean the sentential tune which lacks prosodic focus. Prosodic focus will be discussed in section 4.2.4.) My as yet unquantified observations of many F_0 contours suggest that much of this downtrend can be described as a downstep at each word boundary. That is, the word boundary tone downsteps the succeeding H target. The final constituent in a sentence then usually ends up in a lower pitch range than any other constituent in the sentence. This cues the end of the sentence. On the basis of instrumental evidence, L&I have pointed out the potential for neutralization of word accents in disyllabic words in this position. I&Z have characterized this phenomenon by the phonological rule of L insertion whose effect is to erase the tonal lexical distinctions. However, I show, using minimal pairs, that the lexical tones are still present in this position despite the smaller range for their manifestation (see, sections 4.2.2 and 4.2.4). Therefore, I conclude that a different phonological model is needed from the one I&Z propose. The new model needs to be able to separate the effects of the gradient backdrop pitch trends from categorical tone deletion.

As for sentence-level prosody, words under prosodic focus, in narrow focus utterances, show a higher target for the accent H relative to the same utterance without the prosodic focus. This is true both for the starred tone of the falling accents (H*+L) and the trailing tone of the rising accents (L*+H).

In summary, in this paper I posit three prosodic units for Serbo-Croatian: a phonological word, an intermediate phrase, and an intonational phrase. The declarative sentence pattern of SC shows a continuous alternation between H and L tones. Every phonological word is marked by this pattern, and so is each sentential string. However, the sentence intonation is more than just a concatenation of the word accent tones. The declarative sentence intonation can be accounted for by positing a word-boundary tone, a downstep rule phrase internally, the rule of reduction of pitch range in final position, super H targets for

exceptions. For example, demonstrative pronouns, which function as determiners, thus function words, do bear a pitch-accent. Some conjunctions, such as *pȁ* 'so', *ìako* 'although', *ȁli* 'but', etc. also bear an accent.

discourse-initial segments, and pitch range reduction of post-focus positions. On the basis of contrasts in different melodies, such as declaratives, prompting intonation, and vocative chant, I argue for two different tonal markings of an intonational phrase: two boundary tones, L%, H%, and two phrase accents L- and H- for an intermediate phrase.

The paper consists of three major parts. The first major part, section 3, deals with lexical accents and their properties; the second part, section 4, is concerned with tonal markings of prosodic structure; and section 5 deals with the issues of interaction between the lexical and structural markings. Section 6 concludes by summarizing the proposed analysis of Serbo-Croatian intonation.

2 Methodology

The language that I intend to cover in this paper is the Štokavian–Ekavski variant[2] Standard SC. The analysis presented here is a broad outline investigation. It is based on an instrumental investigation of F_0 contours for close to 300 utterance types, ranging from citation form utterances of single words to three-sentence paragraphs. The intention was to provide a wide coverage of Serbo-Croatian utterance types in order to get an overview of the complete system, as a framework for investigating some specific aspect of the system in a thorough quantitative analysis with careful control of interaction with other sources of systematic variation. This purpose is a result of the need for the more overall picture of the system prior to the later quantitative modelling of specific questions. This is in line with the work done by Pierrehumbert (1980), which provided the groundwork of a complete description of the English intonational system, and which subsequently resulted in the detailed study of pitch range in Liberman & Pierrehumbert (1984). Consequently, results presented here will be more suggestive than quantitative.

All the material uttered by the author was digitally recorded directly into a Sun workstation (Sun4) or Linux box and analyzed using the Entropics Waves program. Materials uttered by the other seven native speakers were recorded in a quiet room on a Marantz taperecorder and then digitized with Waves using a Denon tape player and the Sun workstation. Four of the speakers, including the author, are from Novi Sad, three of the speakers are from Belgrade, and one of them is from Kruševac.

For the purposes of getting an uninterrupted pitch track, almost all of the words and sentences recorded were chosen for their all-sonorant quality. Some exceptions were made when the length or the late position of the accent of the word was crucial in investigating a certain hypothesis and no word with all sonorants was found with those characteristics.

[2]Serbo-Croatian dialects are divided along two parameters: (a) the first parameter is the word for 'what', thus we have *što, ča,* and *kaj* and the corresponding dialects: Štokavian, Čakavian, and Kajkavian; (b) the second parameter is the reflex of the Old Church Slavic vowel *jat*. There are three reflexes of this vowel: [e], [i], and [ije]. Hence the corresponding dialects: Ekavski, Ikavski, and Ijekavski.

Also, as it was important to look at minimal pairs and words with particular syntactic and semantic properties (notably, *wh*-words) it was necessary to include some words that do not have all-sonorant quality.

All the pitch tracks in this paper are utterances performed by the author. This decision is a consequence of the fact that it was not possible to get all the relevant data from all the speakers, and was used to keep the pitch contours consistent throughout the paper for ease of comparison. However, none of the pitch tracks used here for the purpose of illustration are isolated tokens of the type. Pitch tracks were used as evidence only when the same contour ocurred constantly across at least five tokens of the same type of utterance.

3 Lexical Information

3.1 Lexical Accents and Their Distribution

The standard description of the distribution of the accents is that falling accents only occur on the initial syllable and that rising accents occur on any syllable but the last syllable. Thus, rising accents never occur in monosyllabic words since the initial syllable is also the last syllable. So, monosyllabic words necessarily have a falling accent. The traditional way of marking falling accents is: [ˋ] for the short falling, and [ˆ] for the long falling. Some examples of words with these accents and a pitch track of a word under the accent in a sentence medial position are given in the following table.

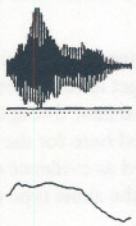
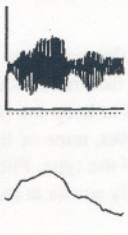
SHORT-FALLING	LONG-FALLING
 jȁlov 'fruitless'	 jâvan 'public'
lȁv 'lion'	lâž 'a lie'
jȁlov 'fruitless'	jâvan 'public'
lȕla 'pipe'	ûlje 'oil'
nȁjava 'announcement'	nâmera 'intention'
nȅminovan 'inevitable'	vôljan 'willing'
nȅravnomeran 'uneven'	vâljan 'rolled'
paradȁajz 'tomato'	ûman 'wise'
ȍrnamentalan	ûmoran 'tired'
ranorȁnilac	rêvija 'review'

Table 1: The F_0 tracks show the two falling accents in words *jȁlov* 'fruitless' and *jâvan* 'public' in a sentence medial position to circumvent discourse or sentence edge effects. The rest of the table provides examples of words under the two falling accents, short and long, with the stress on the first syllable, differing in length.

The traditional way of marking the rising accents is the following: [`] for the short rising, and [´] for the long rising. A rising accent can occur on any syllable but the last and it never occurs on monosyllabic words. Here are some examples:

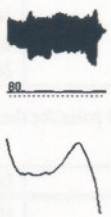
SHORT-RISING	LONG-RISING
	
màlina 'raspberry'	lȉnija 'line'
nàum 'intention'	jáje 'egg'
màlina 'raspberry'	lȉnija 'line'
ȉronija 'irony'	žȉveo 'he lived'
mànjiina 'minority'	aróma 'aroma'
ànemija 'anemia'	nenámeran 'unintentional'
artiljèrija 'artillery	memoári 'memoirs'
neprikosnòven 'sacred'	anulȉrala 'she annulled'
multimiliòner 'multimillionaire'	legitimácija 'ID'
	nacionalizácija 'nationalization'

Table 2: The F_0 tracks show the two rising accents, in words *màlina* 'raspberry' and *lȉnija* 'line', in a sentence medial position to circumvent discourse or sentence edge effects. The rest of the table provides examples of words under the two rising accents, short and long, with varying position of the stress, and differing in length.

3.2 Phonemic Distinctions

Both distinctions, short vs. long and falling vs. rising contrast words. The following sets of minìmal pairs show the contrastive role these properties play:

SHORT-FALLING	LONG-FALLING
rȁd 'eager'	rȃd 'work'
sȁd 'now'	sȃd 'plantation'
ȍran 'plowed' (participle)	ȏran 'disposed'

Table 3: Examples of minimal pairs for the two falling accents.

SHORT-FALLING	SHORT-RISING
mȍli '(s)he begs'	mòli 'Beg!' (imperative)
ȍpasan 'dangerous'	òpasan 'with a belt'
ȍrao 'he plowed'	òrao 'eagle'
ȍsnovan 'founded'	òsnovan 'elementary'

Table 4: Examples minimal pairs for the two short accents.

LONG-FALLING	LONG-RISING
râvan 'plain'	rávan 'flat'
râdi '(s)he works'	Rádi 'to Rada'
nêma 'he doesn't have'	néma 'deaf.fem'

Table 5: Minimal pairs for the two long accents.

SHORT-RISING	LONG-RISING
sèdeti 'to sit'	sédeti 'to go gray'
òpisan 'described'	ópisan 'descriptive'
ràsipan 'wasted'	rásipan 'wasteful'

Table 6: Minimal pairs for the two rising accents.

There are also minimal pairs that cut across both dimensions. That is, words with the same segmental tier but with tonal contrast along both long/short and rising/falling parameter:

LONG-RISING	SHORT-FALLING
(h)rána 'food'	rȁna 'wound'
LONG-FALLING	SHORT-RISING
vâljan 'rolled'	vàljan 'good'

Table 7: Minimal pairs for both duration and pitch oppositions.

3.3 Lexical Tones

Serbo-Croatian pitch-accent can be characterized by the position of stress and the specification of two tone levels, high and low, as already proposed by I&Z in the framework of autosegmental phonology and earlier by Halle (1971). The order and distribution of these

tones relative to the accented syllable corresponds with the type of the lexical accent for which the word is specified.

It has been claimed that the distinction between the rising and the falling accents lies in the fact that the rising accents are bisyllabic whereas falling accents are monosyllabic:

> ' ... relying on perceptual evidence analyzes rising accents as encompassing two syllables ... However, falling accents encompass only one syllable.' (B&M:147 citing Hodge 1958, Bidwell 1963, Masing 1876, Ivić 1958, 1961).

> 'All four accents have traditionally been treated as associated with a single syllable, as the diacritics [...] show. However, only the falling accents are clearly monosyllabic; the rising acccents are disyllabic in nature, as we will see.' (I&Z 1988:227, footnote 2.)

This distinction suggests the assumption that since accent is (by definition) a culminative marker within its domain, the relevant phonetic property should culminate at the accent location; hence a 'pitch accent' should be a pitch culmination, i.e. a peak in the pitch contour localized at the accent. The falling accents are in accordance with this assumption since the characteristic of the falling accents is that the H tone is realized on the accented syllable itself. The rising accents, on the other hand, deviate from this common assumption about alignment between accents and peaks. The H of the rising accents is realized on the post-stressed syllable. This misalignment between the accented syllable and the peak in the rising accents has thus far been couched in terms of durational properties of the accent: monosyllabic, vs. bisyllabic.

Instead of thinking of the two classes of SC pitch-accents, falling vs. rising, in terms of monosyllabic vs. bisyllabic accents, I propose to switch the perspective from the number of syllables necessary to realize the accent peak to thinking of the number of tone targets necessary to realize a rise or a fall, i.e. to consider both of them as being bitonal, where only one of the tones is anchored to a stressed syllable (cf. Bruce 1990). For the falling accents, the anchored tone will be the H, and for the rising accents the anchored tone will be the L. The data show that the second (trailing) tone can be realized on the stressed syllable as well, as in the case of the long-falling accent, but it is usually on the poststressed syllable, as is the case for all other accent types. Consequently, it seems more appropriate to treat the second tone as unassociated rather than anchored to a particular syllable or mora.

This view is more in accordance with the position in Kostić (1983) who argues that all four accents should be treated as bisyllabic. His argument involves the claim that accent peaks are fully realized only in opposition to the following or preceding syllable. However, in order to claim this, he has to exclude monosyllables, which he then treats as exceptions. The position taken in this paper is that pitch accents are only partially linked to a particular

metrical position. Only the first tone of the (bitonal) accent is anchored to the text, and this is the starred tone.

The next three sections explore the consequences of this proposal for the falling accents alone, the rising accents alone, and both together.

3.3.1 Falling Accents

In a declarative utterance, falling accents can be characterized in terms of two tones, H followed by L. Both short-falling and long-falling accents have a H tone on the stressed syllable followed by a L tone. The difference between the two accents seems to be not only in the duration of the syllable under stress but also in the timing of the tonal qualities. In words with long-falling accents the L seems to show up during the stressed syllable whereas in words with the short-falling accent the L starts after the stressed syllable.

Here are some examples of citation forms. Figures 1 through 3 show similar (or minimally contrasting) words, with one, two or three syllables. In all of the figures throughout the paper, the cursors (vertical lines) mark the ends of syllables (or the end of words, when individual syllables are not marked).

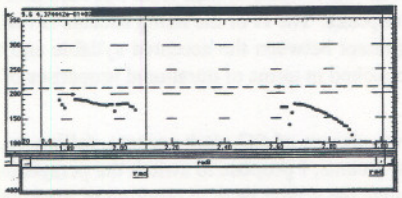


Figure 1 : **Short-falling vs. Long-falling, 1 syllable words.** The first utterance, on the left, is the word: *ȑad* 'eager', the short-falling accent; the second utterance is the word *râd* 'work', the long-falling accent.

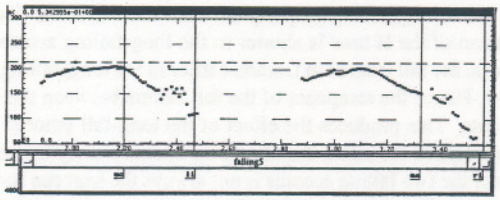


Figure 2 : **Short-falling vs. Long-falling**, two-syllable words, (citation form). On the left is the word *mȍli* 'begs', short-falling accent; on the right is the word *mâri* 'cares', long-falling accent.

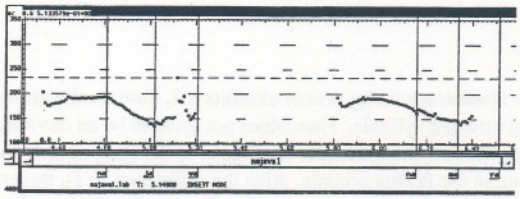


Figure 3 : **Short-falling vs. Long-falling, 3 syllable words.** The first utterance, on the left, is the word: *nȁjava* 'announcement', the short-falling accent. ; the second utterance is the word *nâmera* 'intention', the long-falling accent.

From the three pitch tracks above we can notice that the two accents are associated with a particular shape of F_0. In the long-falling examples, the accented syllable carries both H and L tone, whereas in the short-falling examples, the accented syllable carries only the H tone, and the L tone is realized on the post-stressed syllable. In other words, the alignment of the peak in the long-falling accent is more towards the middle of the syllable, whereas in the short falling-accent it is at the right edge of the syllable.

There is a difference between monosyllables (Figure 1) and disyllabic words (Figure 2), under the short-falling accent. The disyllabic words show the L tone on the post-stressed syllable whereas in the monosyllables the fall is truncated when the word is in isolation.

Despite this difference in monosyllables, the two falling accents are very similar. Consequently, I propose that these accents be represented as H*+L. This representation accurately captures the fact that the H tone is anchored (associated) to the stressed syllable,

whereas the L tone is a trailing tone, which may or may not fall on the stressed syllable, or may even be truncated. The distinguishing property between the two is the duration of the anchored tone. The duration of the H tone is shorter in the long-falling accent than in the short-falling accent. That is, the fall from H to L starts earlier in the long-falling accent than in the short-falling accent. Hence the steepness of the fall differs between the long-falling and the short-falling accents. This produces the effect of the long-fall (shorter H tone) vs. short-fall (longer H tone). Hence the name that they bear seems clearly appropriate. The duration of the vowel under the two falling accents is not always the best cue for which type of accent we are dealing with. The durations of the H tones seem to be more distinct than the durations of the vowels. Although there is a contrast between short and long vowels in unstressed positions, duration is the best cue for stress in SC, as shown by L&I. That is, a short stressed vowel is longer than a short unstressed vowel, and a long stressed vowel is longer than a long unstressed vowel.

3.3.2 Rising Accents

In a declarative utterance a rising accent exhibits a L tone on the stressed syllable and a H tone on the post-stressed syllable. There does not seem to be an obvious qualitative difference in F_0 between the two rising accents analogous to the steepness of the fall or the length of the starred tone in the falling accents. Also the difference in F_0 target is insignificant, when we compare either the peaks or the preceding lows. This is also confirmed by the data reported in L&I (1985). However, there is a difference in vowel quality. The vowels under the long-rise are more peripheral than the vowels under the short-rise. This may be in part due to the difference in duration of the vowel, since the vowel in a word with the long-rising accent is longer than in a word with the short-rising accent. As L&I (1963:93f) report, the long /e/, /o/, and /a/ are more peripheral than their short allophones.

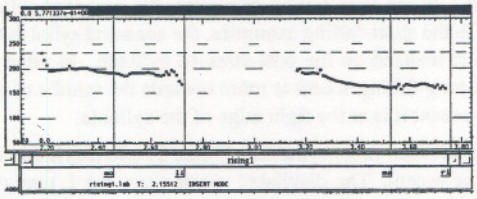


Figure 4: **Short-rising vs. Long-rising**, two-syllable words (citation forms). The first utterance is the word *mȍli* 'beg!' (imperative), short-rising accent; the second utterance is the word *Mári* 'to Mara', long-rising accent.

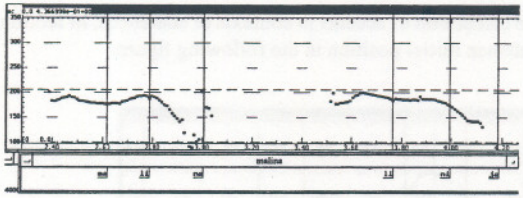


Figure 5: **Short-rising vs. long-rising, 3 syllable words.** The first utterance is the word *málina* 'raspberry', the short-rising accent; the second utterance is the word *lìnija* 'line', the long-rising accent.

In the above figures we can see that the H of the rising accents is not very prominent. This is because these short utterances are citation forms, which inevitably encompasse the phenomenon of final lowering (to be discussed in section 4.2.3). For the purpose of the illustration of this effect I present the word *omalovažàvanje* 'humiliation' in two different enviroments, a citation form and as an initial constituent of a sentence.

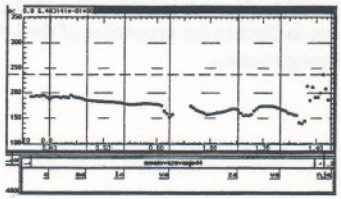


Figure 6: **Short-rising accent.** The utterance is a citation form of the word *omalovažàvanje* 'humiliation'.

This word was chosen for its late accent placement, which allows a long stretch of syllables before the accent. We can notice the L tone, which is anchored to the stressed syllable *-ža-*. The post-stressed syllable is the one that the H tone usually gets realized on. Consequently, the choice for the representation of this type of accent is L*+H. In this case, that is, the citation form, the H tone is affected by the discourse final position, i.e. final lowering. This is the effect that L&I called neutralization of the accents in the final position. However, as will be discussed in section 4.2.2 this effect is due to the pitch-range reduction, and comparison to falling accents clearly shows the preservation of the distinction between the two types of accents. Since utterance final elements are affected by the position, it is especially illuminating to compare the citation form with a non-citation

form. To anticipate the discussion of accents in contexts of sentences, in section 4.2, I show the same word in a sentence initial position in the following figure.

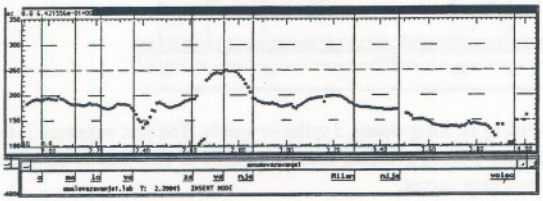


Figure 7: **Short-rising accent.** This is a sentence-initial utterance of the word *omalovažavanje* 'humiliation' ('Humiliation, Milan didn't like.').

As we can see in Figure 7, the H tone of the rising accent is more visible due to the continuation of the utterance. I will return to the sentence level influence on the accents in sections 4 and 5. In the next section I continue to discuss the properties of the lexical tones. The point of interest here are distinctions among accents.

3.3.3 Rising/Falling Opposition

In the previous section the opposition between the short and long accents was shown. The F_0 contour very clearly reflects the opposition between the falling accents. For the rising accents, the F_0 is a less transparent indicator of the contrast between the short and the long rising accent. It is the time course of the H* and the steepness of the fall that create a distinction between the falling accents. The rising accents, on the other hand, do not seem to have as clear a tonal distinction, in terms of the F_0 manifestation: rather they differ in vowel quality.

In this section I present the opposition between the rising and falling accents of the same durational type because this allows us to see the difference between a rise and a fall most clearly, since the length variable is kept constant. Figure 8 shows a minimal pair, the long falling vs. the long rising accent; Figure 9 shows a minimal pair for the short falling/rising opposition.

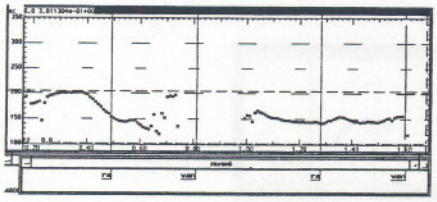


Figure 8: **Long accents.** The first utterance is the word *râvan* 'plain', the long-falling accent; the second utterance is the word *rávan* 'flat', the long-rising accent.

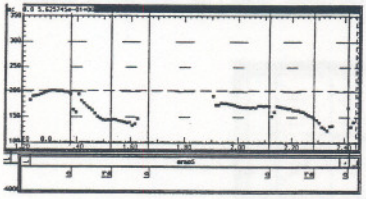


Figure 9: **Short accents.** The first utterance is the word *ȍrao* 'he plowed', the short-falling accent; the second utterance is the word *òrao* 'eagle', the short-rising accent.

The difference between the falling and the rising accents is very clear from the above pitch tracks. The falling accents exhibit a clear fall in the pitch, whereas the rising accents exhibit a small rise or a steady pitch on the post-stressed syllable. The lack of an obvious rise, i.e. a clear manifestation of the H target, in these examples is due to the citation form intonation of the utterances. As we saw in the preceding section, the rising accents do realize the H tone, which is higher from the tone of the stressed syllable, as long as the word is not utterance final. In addition, we can see that the H tone of the falling accents is considerably higher from the H tone of the rising accents. (This is not an artifact of their order in the list since, the reversal of their linear order in production produces the same effect, see Figures 8, and 9.)

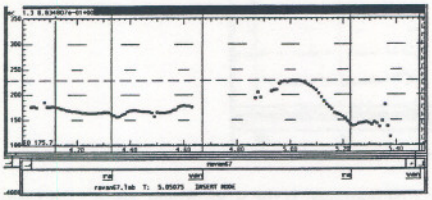


Figure 10: **Long accents.** The first utterance is the word *rávan* 'flat', long-rising accent; the second utterance is the word, the *râvan* 'plain', the long-falling accent.

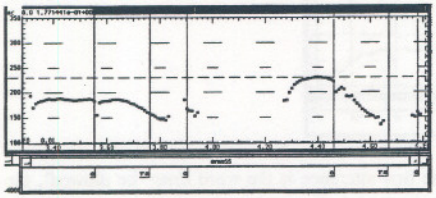


Figure 11: **Short accents.** The first utterance is the word *òrao* 'eagle', the short-rising accent; the second utterance is the word *ȍrao* 'he plowed', the short-falling accent.

This observation has been noted by Kostić (1983) and the F_0 measurements for 3 different pitch ranges of speakers, low, medium and high, from L&I (1963) also support that conclusion.

3.3.4 Lexical Tone Analysis

In this section I give a proposal for analyzing the four lexical tones of the four Serbo-Croatian pitch accents.

As previously mentioned, the falling accents are characterized by the HL melody, whereas the rising accents exhibit LH melody. The tonal distinction on the short-long parameter is manifested with the falling accents (in the steepness of the fall), but not with the rising accents. Schematically, however, the proposal for the four accents can be represented as the following:

	FALLING	RISING
SHORT	σ σ σ | | | μ μ μ | H*+L	σ σ σ | | | μ μ μ | L*+H
LONG	σ σ σ /\ | | μ μ μ μ | H*+L	σ σ σ /\ | | μ μ μ μ | L*+H

Table 8: Surface representation of tones in trisyllabic words with the stress on the first syllable.

In the above graphs the distinctions between falling and rising accents is represented by the HL and LH melodies, whereas the short/long distinction is captured by the mono-moraic vs. bi-moraic status of the syllable to which the accent is associated. So, all the accents are bitonal: however, of the long accent types, only the first tone is anchored to the first mora of the falling accents and the second mora of the rising accents. (The justification for the particular anchoring site within the syllable for the long rising accent requires explanation of one of the phrasal tones and is deferred until section 4.1.)

This differs from the analysis proposed by Inkelas & Zec (1988) (I&Z) in two important ways. First, in their autosegmental analysis, all tone bearing units are specified for tone at the surface. Hence in their theory the structural fact of accent is only a property of an underlying form, whereas in this analysis the accent is viewed as a pitch event localized at the stressed syllable. Second, in I&Z's theory the rising accents are represented as a sequence of two H tones, whereas in this analysis, the rising accents are a LH melody, where the L tone is anchored to the stressed syllable.

The two analysis agree on the representation of the difference between long and short syllables through a moraic structure. For the sake of comparison, I provide a schema of their analysis of trisyllabic words with an accent on the initial syllable after the derivation is completed:

	FALLING	RISING
SHORT	σ σ σ | | | μ μ μ | | | H L L	σ σ σ | | | μ μ μ | | | H H L
LONG	σ σ σ /\ | | μ μ μ μ | | | | H L L L	σ σ σ /\ | | μ μ μ μ | | | | L H H L

Table 9: Predictions of the surface representation of tones in trisyllabic words with the stress on the first syllable, according to Inkelas and Zec (1988). (Compare to Table 8)

Under their analysis, the HL melody of the long-falling accent is realized on the two moras of the accented syllable itself, wheras the short-falling accent is realized across two syllables. In the case of the rising accent, which they represent as HH, the two H tones are associated to the last mora of the accented syllable and the first mora of the post-stressed syllable. This is because in their theory, two adjacent H tones cannot belong to the same syllable. According to my data, both rising accents have the H tone realized on the post-stressed syllable only. That is, the high tone is never realized on the accented syllable or the last mora of the accented syllable. To make the point clearer, I will present the instrumental data of the examples analyzed in their paper, and discuss the predictions their analysis makes about the surface tones.

The following figure shows a pitch track of the two rising accents discussed and analyzed in I&Z:

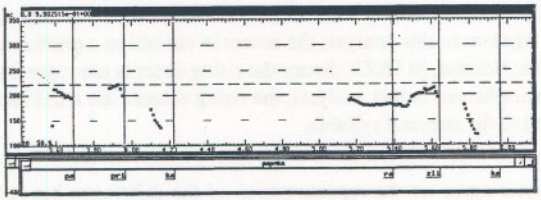


Figure 12: Figure of two rising accents: *pàprika* 'pepper' and *rázlika* 'difference'.

For the purpose of comparing I&Z's analsyis to the one presented here, we must abstract away from the basic difference between these two analyses, such as full vs. sparse surface specification of tones. With that in mind, we can compare only the accent representation in the two approaches.

According to their analysis, the difference between *pàprika* and *rázlika* is in the position of the two H tones: *rázlika* should realize the H tone on the second mora of the first syllable, (*rá-*) and the first mora of the second syllable (*-zli-*), whereas *pàprika* should have a H tone on the first syllable (*pa-*) and the second syllable (*-pri-*). However, we can see that the H tone peak is always realized on the post-stressed syllable. Also, the accented syllable has a L tone in both of these accents. The following pitch tracks also confirm this observation. Figures 13 and 14 show words with long rising and short rising accent, respectively, on the third syllable in a five-syllable word. Both words are uttered in a sentence medial position of a broad focus utterance.

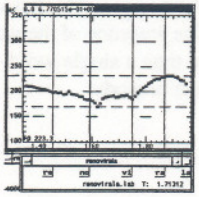


Figure 13: The word *renovírala* 'she renovated' in utterance medial position.

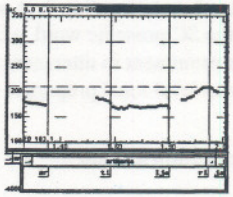


Figure 14: The word *artiljèrija* 'artillery' in utterance medial position.

As Figures 13 and 14 show, the H tone is a property of the post-stressed syllable and it is not shared by the two consecutive moras of the stressed and post-stressed syllable. Thus, I&Z's hypothesis is not consistent with the instrumental data. On the basis of the instrumental evidence both from my corpus and from the corpus presented in L&I, I assume that the H tone of the rising accents is a property of only the post-stressed syllable.

Another clear advantage of assuming that only one tone of these accents is anchored to the accented syllable involves the treatment of monosyllabic words. As we saw earlier, in section 3.1, falling accents can occur on monosyllabic words. If we assume that tones are properties of moras and are anchored in them, then only the long-falling accent would be able to occur on monosyllabic words, since long syllables have two moras and the two tones, H and L would be associated with them. However, the short-falling accent, which is a property of short syllables, thus only one mora is available, would have no place for the L tone. Under my analysis, the fact that both types of falling accents are present in the language falls out as a natural consequence of the fact that only the H tone is anchored to

the stressed syllable. The L tone is capable of being truncated if there is only one syllable for realization of tones.

So, we see that having a representation which anchors only one of the tones to a particular syllable gives us a natural explanation for why monosyllables can have both short and long falling accent. In the theory of I&Z, this fact is not accounted for. However, this same reasoning should give us an explanation for the other part of the distributional fact of the SC accents (that is, an explanation for why rising accents never occur on monosyllables). The traditional explanation has always resorted to the idea that the rising accents are bisyllabic, unlike the falling accents. This is a restatement that still calls for an explanation. But, as I have tried to show in this section: In non-monosyllables, the short-falling accent can also be characterized as bisyllabic. So, is there a natural explanation for the distributional properties of the rising accents within this system? I believe there is. The explanation, offered in the next section, involves reasoning about the functional properties of tones and how densely–distributed similar tones can realize their functions. But, before we can go to that explanation (see section 4.1), it is first necessary to introduce another property of the SC prosodic word, a L word-boundary. The presence of the word boundary tone is more prominent in utterances that consist of more than a single word, thus we turn to the sentence-level tonal properties of the Serbo-Croatian prosody.

4 Structural Information

4.1 Phonological Word

In this section I define the smallest prosodic unit in SC, the phonological word. I show that tonal markings of this prosodic unit are of two types: one demarcative (a left edge tone) and the other culminative (the pitch accent).

4.1.1 Word-Boundary Tone

In addition to the lexical tones considered to be realizations of the word accent type, each phonological word in Serbo-Croatian exhibits a boundary tone as well. That is, each word that bears an accent must have a L boundary tone, which I represent as %L. I will argue that this tone always precedes the lexical tonal realizations for reasons that will be clearer when the discussion of downtrend gets introduced.

In their autosegmental account of Serbo-Croatian tones, I&Z assume that words are specified for the H tones in the lexicon, whereas L tones are assigned late in the process of derivation. For declarative intonation they propose a classical tone association account whereby at the end of a derivation each mora is associated to exactly one tone, either the accent H or the default L. So for example, in a disyllabic word with a long-rising accent

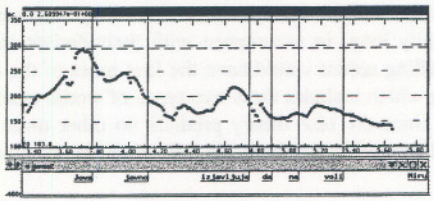


Figure 17: Noun-adverb sequence. The sentence is *Jóva jâvno izjavljuje da ne voli Miru.* 'Jova publically claims that he doesn't like Mira.'

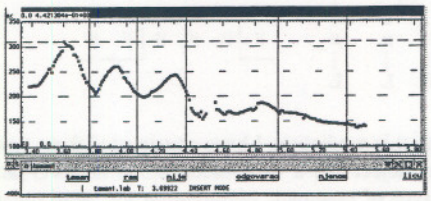


Figure 18: Adjective-noun sequence. The sentence is *Táman râm nije odgovarao njenom licu.* 'A dark frame didn't suit her face.'

In all three pitch tracks (Figures 16–18), the first word is disyllabic and has a long rising accent on the first syllable and the second word has a long falling accent on the first syllable. Since the first word is disyllabic, we know that the H tone will be realized on the second (i.e. final) syllable. The second word, having the falling accent on the first syllable must exhibit a H tone on the first syllable. If there were no word boundary tones, simple concatenation of these two words should produce a steady pitch line representing the two H tones, one from the final syllable of the first word and one from the initial syllable of the second word.

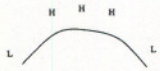


Figure 19: Schematic representation of the prediction for two consecutive H tones according to I&Z's theory.

However, as we see in Figures 16–18, the two H tones are separated by a dip in pitch. This intervening valley I take to be the evidence for the %L boundary tone.

We may ask where the %L tone belongs. Does it belong at the end of a word? or is it the initial leading tone of every word, i.e., the beginning of every phonological word? For reasons that have to do with overall declination pattern, and patterns in sentence initial and final positions, I will assume that the word boundary tone is at the beginning of the word. I will argue for this hypothesis in section 4.2.2, where I discuss utterance final position.

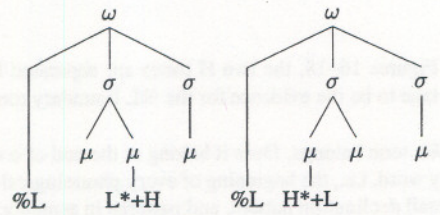
The acceptance of the L word boundary tone then gives us the following picture of the word tones for trisyllabic words with an accent on the first syllable.

	FALLING	RISING
SHORT	σ σ σ | | | μ μ μ | %L H*+L	σ σ σ | | | μ μ μ | %L L*+H
LONG	σ σ σ /\ | | μ μ μ μ | %L H*+L	σ σ σ /\ | | μ μ μ μ | %L L*+H

Table 10: Surface representation of tones in trisyllabic words with the stress on the first syllable and including the initial word boundary tone.

In terms of the theory of tone association to the prosodic hierarchy proposed in Pierrehumbert and Beckman (1988), the boundary tone is associated to the word node, whereas the lexical tones are associated to the stressed syllable. So, the tone structure of the sequence of the first two consecutive words depicted in Figures 16–18 can be represented as follows:

(2)



We are now ready to get back to the question we left at the end of the previous section regarding a possible explanation for the distribution of the rising accents. To review, we said that assuming that the accents in SC are bitonal, where only one of the tones is associated to the stressed syllable, the model requires no extra mechanism to explain the ocurrence of the falling accents in monosyllables, as I&Z's theory would certainly require. The question that we could not answer at the time concerned the curious distribution of the rising accents: they never appear on monosyllables or on the last syllable in a word. Positing a word boundary tone at the beginning of the word creates the following sequence for the rising accents: %LL*+H. The %L tone serves the delimitative function, whereas the L*+H (the pitch accent) serves the culminative and the contrastive function — the accent is rising not falling. The sequences %LL*+H, %LH*+L, and two durations would be hard to contrast on a single syllable. That is, I propose that it is the initial word boundary tone which creates an impossible sequence for monosyllables under the rising accents due to crowding of tones of the same type (i.e. L tone) with different functions, particularly when the duration of the starred tone needs to separate long accents from short ones. If rising accents did occur on monosyllables then we would need to be able to make a four-way distinction in the timing of the rise on a single syllable. At this point this is a very speculative statement and more research would be needed to confirm this hypothesis.

It is worth pointing out a historical perspective on the distribution of accents. The synchronic situation is a product of the so-called Neoštokavian stress shift (started in the 15th century). There were only the two falling accents in the old Štokavian dialects. The retraction of the stress from the syllable associated with the H tone to the preceding syllable gave rise to the rising accents. In other words, the rising accents are the reanalysis of the situation that arose when the stressed syllable was no longer associated with the H tone. This separation of the link between a stressed syllable and a H tone thus seems to be adequately captured in the proposal given in this paper.

4.1.2 Double-Accented Words

An additional piece of evidence for the %L word boundary can be found in double-accented words. The concept of a doubly accented word may seem odd since I am assuming

that a definition of a phonological word is a prosodic unit with only one pitch-accent and, as I am also arguing here, a word boundary tone. However, there seem to be exceptions to my definition of phonological word. It is possible to find examples of double-accented phonological words. These words are always polymorphemic, and are in free variation with variants realized with one pitch accent. They give us an example of what a string of pitch accents looks like without a word boundary.

The following pitch track shows two near-identical sentences containing a word *najmanja* 'the smallest', which can have either the long-falling accent on the first syllable *naj-* or it can have two long-falling accents on the first and on the second syllable.[3] The utterance on the left side contains the one-accented version and the utterance on the right side, the two-accented version.

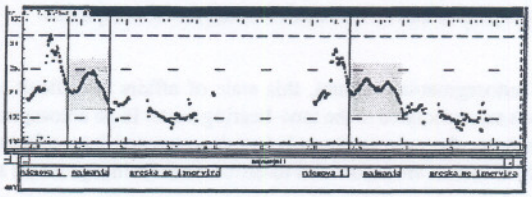


Figure 20: Two utterances of the sentence: *Njegova i najmanja greška me iznervira* '(Even) his smallest mistake irritates me.' The shaded parts of the f0 represents the word *najmanja* in the two utterances; the one on the right contains a double-accented word.

If there were a word boundary tone in the double-accented word, then the L tone between the two peaks would have been lower and the two peaks would not have been of the same height (as is the case for succeeding words, since downstep is a part of every intonational phrase, and will be discussed in section 4.2.3). To see the difference between a double accented word and two words under the same accent as the double accented word, I provide the following pitch track, where the word *najmanja* occurs in the first utterance and the words *moj mali*, which have the same accent pattern across the same number of syllables, are in the second utterance.

[3]Stevanović (1989:431) notes that some long forms of superlatives obligatorily have two accents, such as *najdostojanstveniji* 'the most dignified'.

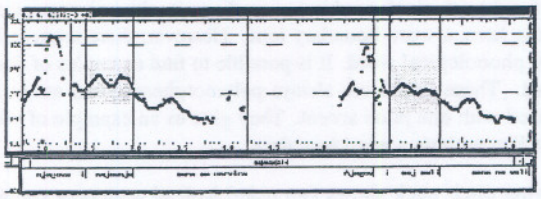


Figure 21: The first utterance, on the left, *Njegova i najmanja greška me iznervira* '(Even) his smallest mistake irritates me.' contains a double accented word; the second utterance, on the right, *Njegov i moj mali mene ne voli* 'His and my young one doesn't like me.' contains two words, with the same number of syllables as the first utterance, in the position of the double accented word. The shaded parts of the f0 point to the parts in the two utterances relevant for the comparison.

In a purely autosegmental account, this state of affairs is difficult to account for since all surface tones are associated to the tone-bearing units. In an autosegmental/metrical account, argued for here, the tone string is being decomposed into culminative and demarcative tones, which in turn are associated to different units in a prosodic structure.

To summarize, I have introduced a new concept into the description of the SC prosody, the %L word boundary tone. The evidence presented so far for the word boundary comes from two sources: the pitch dip observed in sequence of words under a rising and a falling accent, and the pitch level differences observed when this dip is compared to the dip in sequences of falling accents in polymorphemic words such as *najmanja*. These differences in F_0 pattern within a morphological word and across two words is easily explained in autosegmental/metrical account. The two pitch contours can be given two different parses by having the two L tones be part of different constituents in prosodic structure. In more classical autosegmental accounts with only one type of tone-bearing unit, on the other hand, both strings are analyzed as the same HLH sequence.

The third piece of evidence for a L word boundary tone will be introduced in section 4.2.3, where I will try to argue for the downstep model of the downtrend in SC. If the idea that downstep is a consequence of the alternation between H and L tones is correct, as suggested in autosegmental literature on African tone languages (see e.g. Clements and Ford 1979, 1981) then we might expect SC to use the H L alternation as a trigger for downstep regardless of where the L comes from in the grammar of tone. But, as can already be seen in Figure 20, the sequence of two peaks in double-accented word has the peaks at the same level, whereas the sequence of two peaks in two consecutive words, which I claim are separated by a %L word boundary tone, the two peaks are not at the same level. Consequently, it seems reasonable to speculate that it is the presence of a %L

word boundary that would account for the downstep model in SC very naturally, rather than saying that downstep occurs with any H L sequence. But before we can accept that as evidence, the nature of the downtrend needs to be examined in more detail to see whether a downstep account is tenable.

In this section, I also hypothesized that the presence of the %L boundary tone allows us to explain the distribution of the rising accents by assuming that the sequence of %LL* tones followed by +H needs more space for realization than a single syllable.

4.1.3 Clitics

Serbo-Croatian has morphological words which lack both stress and accent and thus are called toneless words, i.e., clitics. These words are prosodically dependent on phonological words. There are two types of clitics in SC, proclitics and enclitics. Prepositions are proclitics, they cliticize to the noun that follows them.[4] Short forms of personal pronouns and verbal auxiliaries are encltics. They are the so-called second position clitics, they cliticize to the preceding word. In this section, I show that proclitics and enclitics differ not only with respect to whether they precede or follow their host but also in their tonal specification. I argue that proclitics realize an edge tone, whereas enclitics have no tonal properties.

The shaded parts in Figures 22–24 show the prosodic behavior of a preposition (*prema* 'towards'), which is a proclitic, in three different positions in the sentence, initial, early medial, and late medial positions. Absolute final position of a preposition is not possible, since preposition stranding is not a syntactic option.

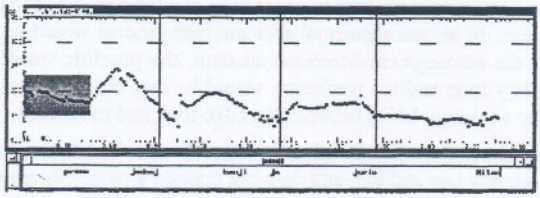


Figure 22: *Prema jodnoj banji je jurio Milan* 'Milan was rushing towards the iodine spa.'

[4]Negative particle *ne* is also a proclitic, however I will exclude it from consideration in this paper. It cliticizes on to the verb that it modifies. Sometimes it even incorporates into the verb, i.e., *nísam* < *ne jesam* 'am.not'. When unincorporated, it can sometimes attract the accent *né znam* 'I don't know'. Prepositions do not attract the accent. It thus differs from prepositions, in the dialect I am describing.

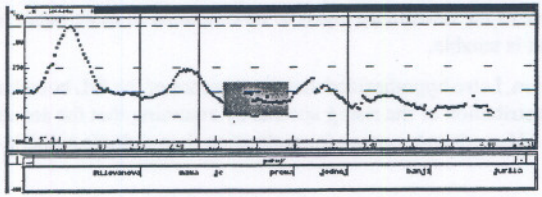


Figure 23: *Milovanova mama je prema jodnoj banji jurila* 'Milovan's mother was rushing towards the iodine spa.'

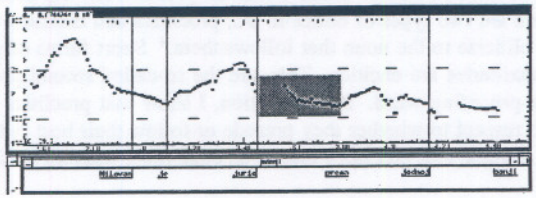


Figure 24: *Milovan je jurio prema jodnoj banji* 'Milovan was rushing towards the iodine spa.'

The sequence of a proclitic and its host, *prema jodnoj* 'towards iodine', in Figures 22–24, according to a traditional wisdom is a sequence of a toneless word and a word with the long-falling accent. In an autosegmental account, these moras would be assigned a L tone by default. In the autosegmental/metrical account, the proclitic would be realizing the left word boundary tone and the prediction would be that the F_0 associated with the preposition would be an interpolation between the edge tone and the accent.

However, the above three figures allow us to see that the F_0 of proclitics in all three positions is comparatively low and flat, and does not contain a peak. Moreover, the rise to the peak of the falling accent does not start until the beginning of the word that bears that accent. That is, the F_0 stretch relating to the preposition seem clearly separated from the F_0 relating to the host of the preposition. I propose that we analyze proclitics as a sequence that realizes a left edge tone, an initial %L word boundary tone. That is, proclitics add edges with no heads. Thus, a sequence of a proclitic and its host is a realization of two edge tones and an accent: %L %L T*+T. The motivation for this analysis comes from the F_0 on the proclitic, which starts low and stays low (or even falls slightly) until the beginning of the word the clitic is attached to. Since, in examples like above, we have a word under the long

falling accent, i.e. H*+L, without a %L at the host's edge we would predict a steady rise towards the accent H tone from the left edge of the proclitic. However, we always get a steep rise only at the beginning of the left edge of the host of the proclitic and not from the left edge of the proclitic itself.

In contrast to proclitics, enclitics do not have an edge tone associtated with them. They are truly toneless morphemes. In SC, enclitics cluster in the so-called second position. The second position is an elusive concept because its best definition is a disjunction: 'the second position is either after the first accented word, or after the first accented constituent' (see Browne 1967:5, who was the first to discuss SC enclitic placement in the generative literature).

In Figures 22–24, we have an auxiliary clitic *je* occuring in various positions. In Figure 22 it occurs after the third phonological word, in Figure 23 after the second, and in Figure 24 after the first. In all of these figures we can observe that the clitic functions as material that interpolates between two tonal specifications: the accent of its host and the %L word boundary tone of the succeeding word.

To summarize, proclitics and encltics differ prosodically. Proclitics carry a word edge tone, whereas clitics do not.

In this section I have argued, on the basis of tonal evidence, for a prosodic unit which I call the phonological word. This prosodic domain is defined by a word boundary tone as a delimitative marker and a pitch accent as a culminative marker. I have also shown that in some cases we have a unit which may lack a culminative marker, such as proclitics, or a unit which may lack a delimitative marker, such as double-accented words. These prosodic units are fused with other units that complement them to form a phonological word. It would be interesting to see if tonal evidence for this prosodic unit can be strengthened by segmental evidence as well.

4.2 Intonational Phrase

In this section I discuss a prosodic constituent higher than the phonological word, namely, the intonational phrase. Two major properties of this prosodic constituent are phrase accents, boundary tones, and pitch range manipulation. That is, pitch range expansion and contraction, and boundary tones can be used as probe for prosodic structure above the word in SC. I show that this prosodic constituent realizes four types of tones: two boundary tones, L% and H% in combination with two phrase accents, L- and H-.

4.2.1 Initial Position

Both the sentence initial position and the discourse initial position in an utterance have the highest H target of all the phonological words in a sentence. However, the two

differ by the level of H. The utterance inital H is higher than the sentence inital H. This position is set off from the rest of the words by the relatively higher pitch target regardless of the syntactic status of the constituent or the word. That is, the H tone in the first position is higher than the H in the second position regardless of whether the phonological word is a syntactic unit by itself or a part of a larger phrase.

To illustrate this point, consider a more elaborate utterance (in Figure 25) consisting of three sentences instead of just one. We can notice that the H in each subsequent sentence initial position is slightly lower than the preceding one. Thus, the absolute utterance-initial position is always set off from all the others by its highest H target. We can see this clearly in the pitch track in Figure 23, representing the following text:

(3) a. Milovanova mama je žurila na voz.
Molovan's mother aux hurried on train
Milovan's mother was rushing to catch a train.

b. Nije imala vremena da gleda ljude u prolazu,
not.aux had time that look.at people in transit
She didn't have the time to observe people around her,

c. ali je njenu pažnju Marija ipak privukla.
but aux her attention Mary still attracted
but Mary still managed to attract her attention.

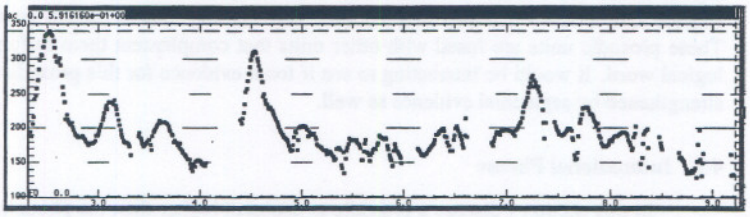


Figure 25: Three consecutive sentences from example (3), showing the set off peaks of the inital constituents and scaling of the three peaks across discourse.

Each pair of adjacent sentences in the above sequence is separated by a short pause, yet their initial H targets create an internal slope thereby bringing cohesiveness to the whole utterance. The internal structure of the three sentence utterance is reminiscent of English utterances as documented by Lehiste (1975). Lehiste showed that in English paragraphs, utterances are characterized by a certain intonation structure, the so-called 'paragraph intonation'. The relationship between pitch range and discourse topic structure has also been

suggested by Brown et al. (1980) and Hirschberg & Pierrehumbert (1986) for English, and by Grønnum (1985) for Danish.

4.2.2 Declarative Sentence Final Position

The sentence final position is also characterized by its distinctive intonational shape. Any type of a syntactic constituent with any type of a word accent in a sentence final position shows a highly reduced pitch range with the pitch very close to the speaker's base line. This effect is treated as final lowering in Inkelas and Zec (1988:240) or laryngealization by Lehiste and Ivić (1986:186). L&I point out that the effects of laryngealization very often seem to lead to neutralization of the accents in a sentence final position. This leads I&Z to posit the rule of final lowering, which stipulates the insertion of a L tone on the last syllable of the last word over-riding the H of the lexical accent (which in their model is always an associated tone). This rule makes a prediction that accents in disyllabic words are neutralized in sentence final position. The data that I have collected show that the distinctions among the word accents are still preserved (Godjevac 1999). However, the distinctions are reduced relative to the initial or medial positions in a sentence of this type. Hence, I would argue that a phonological representation should not include a rule like I&Z's final lowering, since the phenomenon appears to be an effect of some aspect of backdrop pitch range, which Figure 25 shows can be varied in continuous but systematic way to gradiently signal a position within the larger discourse.

The following two figures show the difference induced by the sentence position on the same words. In the first figure, Figure 26, we can see the initial position of the word *mlâda* 'young' and the final position occupied by the other member of this minimal pair, the word *mlâda* 'bride'. In the second figure, Figure 27, the two words are in the reversed positions. This illustration allows us to see the difference between a falling accent and a rising accent in the sentence initial vs. final position.

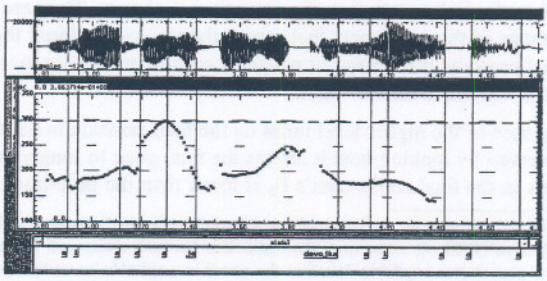


Figure 26: *Mlàda je devojka mlâda* 'A/The young girl is a/the bride.'

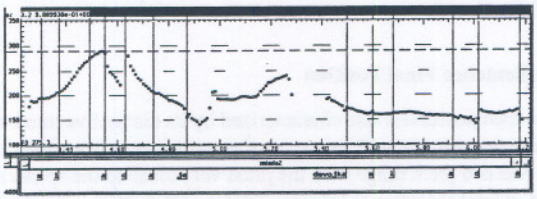


Figure 27: *Mlâda je devojka mlàda* 'A/The bride is a young girl.'

From the above figures we can see that the rising accent stays level in the final position, whereas the falling accent is falling, and it actually becomes laryngealized. Laryngealization is a low pitch common to final falling intonation. Note that in Figure 26 vocalization seen in the wave form continued past the point where the pitch extraction algorithm gives up. The wave form shows the irregular (laryngealized) pulses whereas the pitch track is empty. Therefore, there is a clear differentiation between the two accents even in the sentence final position. The reduction of the pitch range did not erase the lexical tonal distinctions.[5]

The rule of final lowering of I&Z is an insertion of a L tone on the final mora of an utterance. This rule predicts that the final syllable of *mlâda* should be lower than the last mora of the first syllable, which would be assigned the lexical H. As we can see from the Figure 27, that prediction is not borne out.

Instead of positing a final L insertion rule, which effectively erases the lexical H, I posit a L- phrase accent. That is, declarative utterances are marked by a L- phrase accent, followed by a L% boundary tone. The accent and the boundary tone are properties of a higher level phonological constituent, the intonational phrase. They are realized by lowering the pitch range of the constituent that carries the phrasal marking: the right-most constituent in neutral prosodic conditions, or whatever constituent is chosen in the case of prosodic focus, as we will see in section 4.2.4.

The consequence of the higher level tones on the final position in neutral prosodic contexts can be observed by looking how it affects the final peak in longer utterances. It is clear that the peak in the final constituent's F_0 is lower than the proportional reduction

[5]There is some additional evidence for the preservation of the falling/rising distinction. In her acquisition study of SC accents, Kariya (1983:60) notes that 'the distinction between rising and falling accents was evident from patterns of post-stressed syllable deletion: the vowel in a syllable immediately after a falling accent was much more likely to be whispered or deleted than the vowel in a syllable immediately after a rising accent.

based on the preceeding peaks would have predicted. Schematically, we could represent that relationship in the following way:

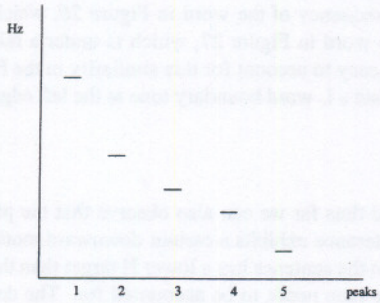


Figure 28: Schematic representation of the peak-proportions for an utterance of 5 phonological words.

I claim that this is a direct consequence of the two final L tones, L-L%, associated with the higher level prosodic constituent, the intonational phrase.[6]

Thus, what seemed like a conspiracy against lexical accents in final position is just a consequence of tonal marking of higher level prosodic constituents. Lexical accents are still present in the final position, but they are affected by the higher level tones. This analysis predicts that the shorter the content word in the final position, the more crowded tones will be, and consquently the more difficult it would be to see them by observing (measuring) the F_0. Hence, under the assumption that tones are only properties of syllables, the conclusion that the accents are neutralized in this position seemed inevitable.

There is another piece of evidence that accents are not neutralized in the final position: they show up clearly under prosodic narrow focus. I will present this evidence in section 4.2.4, as a part of the discussion of prosodic focus.

Before I close this section I want to bring up again the question of where the L word boundary tone belongs. I proposed earlier that we assume that the word boundary tone belongs to the left edge, that is, at the beginning of every word. My reasoning for this has to with the intonational phrase initial words. Since falling accents are specified for a

[6]This property may be similar to what Liberman & Pierrehumbert (1984) found for English and for which they proposed a phonetic rule of final lowering. If under a more scrupulous investigation the sequence of the two L tones cannot account for the F_0 in the final position, a rule of final lowering analogous to the rule proposed for English would also be necessary for Serbo-Croatian.

HL melody we would expect that they would start higher than words with rising accents which have a LH melody. However, if we compare the initial constituents in Figure 26 and Figure 27 (which are minimal pairs), we see that both start with the same F_0, which is relatively low. The fundamental frequency of the word in Figure 26, which is under a rising accent, stays low, whereas the word in Figure 27, which is under a falling accent, rises steeply to reach its H tone. It is easy to account for this similarity in the F_0 pattern of the beginnings of words, if we postulate a L word boundary tone at the left edge of a word.

4.2.3 Downtrend

From the all figures presented thus far we can also observe that the pitch contour of the SC declarative, broad focus utterance exhibits a certain downward motion. That is, each subsequent phonological word in the sentence has a lower H target than the preceding one. This behavior of the declarative tune needs to be accounted for. The decline in the pitch level as a declarative utterance evolves seems to be a fairly common phenomenon crosslinguistically (Ladd 1996:73).

Modelling of the pitch decline across an utterance can be done in several ways. One model reduces the high and the low tones in a declination mode equally, keeping the tonal space the same over time. A different model reduces high tones (Pierrehumbert & Beckman 1988). A more complicated model reduces both high and low tones but each of them differently (Pierrehumbert 1980). A schematic representation of these models is illustrated in the following figure.

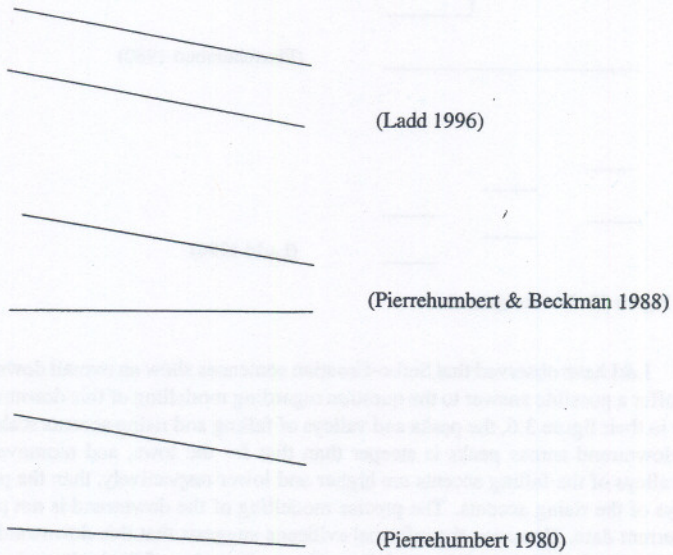
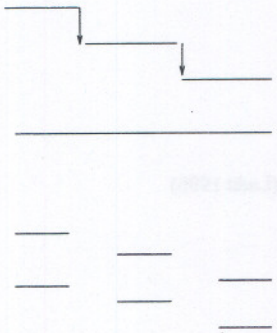


Figure 29: Declination models

However, declination is not the only way to account for the pitch decline across an utterance. Pierrehumbert (1980) has shown that for English, it is also possible to assume a downstep model. The difference between a downstep model and a declination model is in the predictions of the way pitch level is realized between relevant peaks. According to a declination model, the pitch level declines all the time, that is, even between the relevant peaks. According to a downstep model, the pitch level is level between the relevant peaks and it only declines in a step motion at a relevant point. Downstep models can also be modelled in several ways, similarly to the declination models mentioned above. A schematic illustration of the downstep models is presented in the following figure.



(Pierrehumbert 1980)

(Ladd 1996)

Figure 30: Downstep models

L&I have observed that Serbo-Croatian sentences show an overall downtrend. Their data offer a possible answer to the question regarding modelling of this downtrend. As they show in their figure 3.6, the peaks and valleys of falling and rising accents scale differently. The downtrend across peaks is steeper than that for the lows, and moreover, the peaks and valleys of the falling accents are higher and lower respectively, than the peaks and the valleys of the rising accents. The precise modelling of the downtrend is not possible with the current data. However, the informal evidence suggests that this downward slope is not a continuous declination of the pitch but rather a downstep of the highs and possibly lows within an intonational phrase. In addition to the above mentioned models of downtrend, it is possible to imagine a model that may involve a combination of a downstep of the highs and a declination of the lows. In what follows, I will show what kind of evidence we have and will tentatively argue for a downstep model, although precisely which type of a downstep model will be left open.

A typical effect we find in connection to downtrend in SC can be seen clearly in Figure 31.

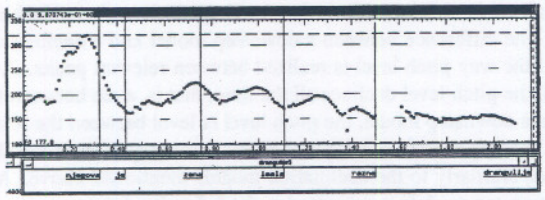


Figure 31: *Njegova žena je imala razne drangulije* 'His wife had all sorts of junk.'

In Figure 31, we can notice that the H of every word (i.e. the lexical H tone, which is part of every phonological word regardless of the accent) is slightly lower than the preceding one, modulo the first and the final words, which seem to be subject to their special position in an utterance. Thus, there is a clear effect of the downtrend in a declarative utterance. If we assume that the pitch range falls steadily throughout the utterance, as in a declination model, then this is not surprising. What a declination model also predicts is the steady decline of the pitch even in syllables that are marked for the same tone. So, a good testing ground for this prediction would be an utterance consisting of longer words whose accent is later in the word. A good candidate for this would be the word *omalovažávanje* 'humiliation'.

The following three figures show the word *omalovažávanje* 'humiliation' in the three sentence positions, initial, medial and final, respectively.

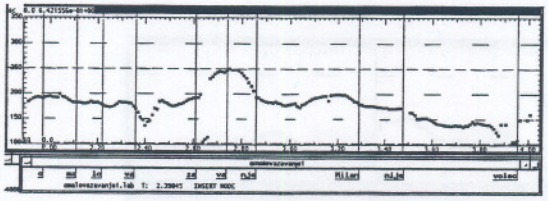


Figure 32: *Omalovažavanje Milan nije voleo* 'Humiliation, Milan didn't like.'

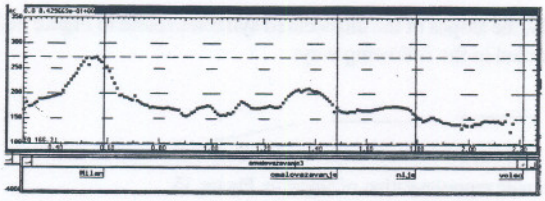


Figure 33: *Milan omalovažavanje nije voleo* 'Milan didn't like humiliation.'

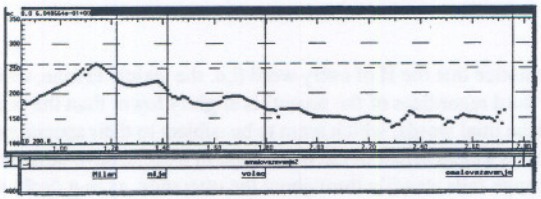


Figure 34: *Milan nije voleo omalovažavanje* 'Milan didn't like humiliation.'

The pitch tracks in Figures 32–34 show us that the syllables which are not affected by the lexical accent or the boundary tone in sentence initial and sentence final positions do not stay level completely, but seem to show a slight slope, whereas in medial position they stay level. This is even more prominent in Figure 35 in which the relevant portions are shaded:

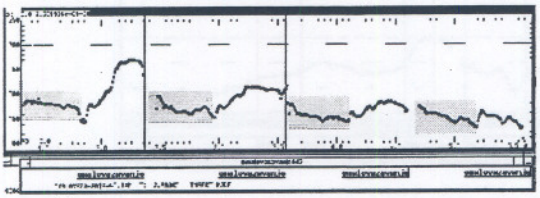


Figure 35: *Omalovažavanje, omalovažavanje, omalovažavanje, omalovažavanje* 'Humiliation, humiliation, humiliation, humiliation.'

Schematically, the slopes of the unaccented syllables found in Figure 36 (the shaded areas) can be represented in the following way:

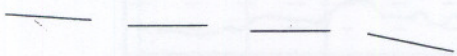


Figure 36: Schematic represenatation of slopes in Figure 35.

Figure 36 schematically represents the slight slope of the unaccented syllable strings in initial and final positions in a sentence. We do not observe the same effect in medial positions. Since a declination model would predict a slope in medial positions, I

propose a downstep model as an account of the downtrend within an intonational phrase.[7] Although these stretches may seem to be too short for a definite conclusion regarding the L tones, we may confidently say that L tones are not subject to whatever it is that is reducing the successive H tones to the same degree. Sentence initial position and sentence final position would have to be accounted separately. However, given that these two positions have additional properties not shared by others (discussed in section 4.2.1 & 4.2.2), they require a special treatment anyway.

I want to introduce another property of the SC prosodic system that I will call a 'pleating effect', which seems to be a direct function of the length of the utterance and is relevant for any modeling of the downtrend. To my knowledge, this was first discussed in Kostić (1983). Basically, the pitch range gets partially reset to a higher target at constituent boundaries as the utterance gets longer.[8] This effect has also been noted for Japanese by Kubozono (1992), although he called it 'metrical boost' and gave it a specifically rhythmic interpretation. As he explains, the phenomenon:

> '... can be understood [in such a way] that the downstepped phrase has been raised by the phonetic realization rule of metrical boost to such an extent that it is now realized higher than the [previous] phrase. This case is typical ... at major syntactic boundaries ...'

I will illustrate this phenomenon in SC by a series of three pitch tracks that represent a successive lengthening of a simple sentence. The three sentences are as follows:

(4) a. Njegova žena je imala dve violine.
his.NOM wife.NOM AUX had two violins.ACC
'His wife had two violins.'

b. Njegova žena je imala dve violine iz istog perioda.
his.NOM wife.NOM AUX had two violins.ACC from same period
'His wife had two violins from the same period.'

c. Njegova žena iz prvog braka je imala dve violine iz istog perioda.
his.NOM wife.NOM from first marriage AUX had two violins.ACC from same period
'His wife from his first marriage had two violins from the same period.'

An utterance of the sentence in example (4a) has no pleating effect, as the following pitch track shows:

[7]The slight difference in the medial stretches defined by L tones can be explained by treating the L word-edge tone and the L* of the accent as different targets for the L.

[8]The partial reset of the declinaton was discussed in Ladd (1984, 1988); however the partial reset was a function of scope disambiguation between two conjunctions, 'and' and 'but'. The partial reset may have the same function in SC as well, but, it need not, as in the case I am presenting. It can simply be a function of the length.

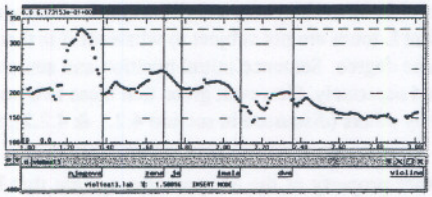


Figure 37: *Njegova žena je imala dve violine* 'His wife had two violins.'

As we can see, the H targets get lower and lower in the utterance as we proceed from the beginning to the end. The next two pitch tracks illustrate a 'pleating effect'.

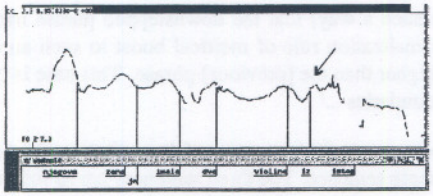


Figure 38: *Njegova žena je imala dve violine iz istog perioda*'His wife had two violins from the same period.'

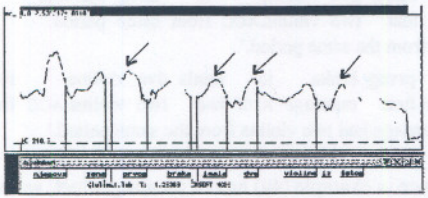


Figure 39: *Njegova žena iz prvog braka je imala dve violine iz istog perioda* 'His wife from his first marriage had two violins from the same period.'

The pitch range reset at each phrase is done in such a way that the level of the H tone is reset to the same level, or a slightly higher level than the preceding H tone, thereby

breaking up the downtrend. This is the 'pleating' effect. There is no focal prominence on any of these constituents in which the first H was reset to a higher pitch range.

Beckman & Pierrehumbert (1986) show, the reset of the pitch range can serve as evidence for a phrase boundary, as they used it, in addition to pausing, in arguing for an intermediate phrase in English. So, is the 'pleating effect' then a matter of prosodic phrasing? That is, do the points of reset correspond to prosodic boundaries of any sort? And if so, what type of prosodic constituent do these points correspond to? I propose that these units are phrases which are the domain for the local manipulation of the pitch range, i.e. downstep, as also proposed for English by Pierrehumbert (1980). I will call them intermediate phrases.

Formation of intermediate phrases is sensitive to syntactic boundaries. However, the syntactic boundaries that seem to be relevant do not form a natural class. Consider the resetting of the pitch range in Figures 38 and 39. In Figure 38, the reset is done at a boundary between a noun phrase and its PP modifier (adjunct). In Figure 39, there are four reset points: (1) at the same point as in Figure 38, (2) at the point of a syntactic head/complement boundary (V and NP), (3) at the point between the last constituent in the subject NP and the first constituent of the VP (i.e., the main verb), and (4) at the boundary between an NP and its PP modifier (the same boundary as in (1)). These are the two basic types of syntactic boundaries: head/complement and head/modifier. Because both types of syntactic boundaries can function as reset points, I take this to be evidence that intermediate phrases cannot be derived by an algorithm sensitive to syntactic relations, such as the one proposed by Nespor & Vogel (1986). In addition, an end-based algorithm, as proposed by Selkirk (1986), also does not make the correct prediction. An end-based algorithm would predict intermediate phrases in shorter utterances, such as those depicted in Figure 37, where we never find them.

As I have shown, and as Kostić (1983) has also claimed, intermediate phrase formation is a function of the length of an utterance. Kostić claims that the relevant crossover point is 5 words. That is, utterances that are longer than five words will inevitably be realized as more than one grouping of words, or in our terminology more than one intermediate phrase. However, how many intermediate phrases an utterance of six words will have is not determined. As Kostić argues, there could be two or three. That is, we expect speakers to differ in the way they chunk the utterance. Thus, even though syntactic boundaries are relevant to the formation of intermediate phrases, knowing where the syntactic boundaries are will not necessarily give us the correct grouping of words into intermediate phrases, because they differ both within speakers and across speakers.

4.2.4 Prosodic Focus

So far, we have looked at utterances which do not have any prosodic prominence except for the lexical stress. That is, prosodically they are all broad focus. Semantically, however, the focus domain is determined by the interaction of this prosodic property and word order considerations. So, prosodic broad focus is what I call neutral intonation.

In this section I turn to prosodic focus. By prosodic focus, I mean prosodically marked emphasis on some constituent in a sentence. Serbo-Croatian allows its constituents to be prosodically focused, which in turn signals semantic focus as well. Since semantic focus is crucial for interpretation of utterances, both for their truth conditional and non-truthconditional meaning, the investigation of prosodic focus is crucial in the overall understanding of the language. Semantic focus in SC can be signaled via word order as well as prosodically. For word order to function as semantic focus marking, prosodic focus must be absent. That is, the sentence intonation must be neutral.

Any phonological word (words that can bear accent) can be prosodically focused regardless of its position in the sentence and its syntactic function. The phonetic effects of prosodic focusing are pitch range manipulation of the focal constituent and its environment. A focal constituent is realized in a slightly expanded pitch range, whereas post-focal constituents are realized in a significantly reduced pitch range. In addition, pre-focal constituents may also be affected by a slight compression of the pitch range. The following five figures show the same sentence with different prosodic focus patterns. The first figure shows the sentence *Jelena daje Mariji limun* 'Jelena is giving Mary a lemon' in a broad focus utterance. The next four figures show the same sentence with a prosodic narrow focus on one of the constituents in the sentence, a different one in each case.

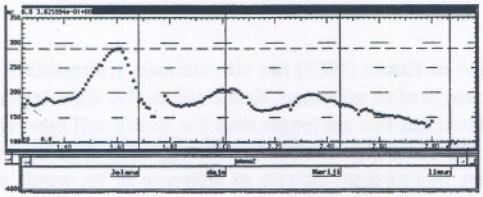


Figure 40: *Jelena daje Mariji limun* 'Jelena is giving Mary a lemon.'

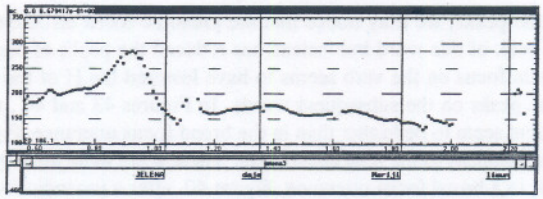


Figure 41: *Jelena daje Mariji limun* 'JELENA is giving Mary a lemon.'

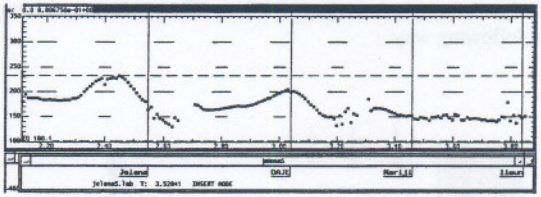


Figure 42: *Jelena daje Mariji limun* 'Jelena is GIVING Mary a lemon.'

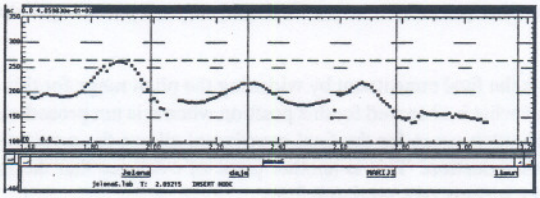


Figure 43: *Jelena daje Mariji limun* 'Jelena is giving MARY a lemon.'

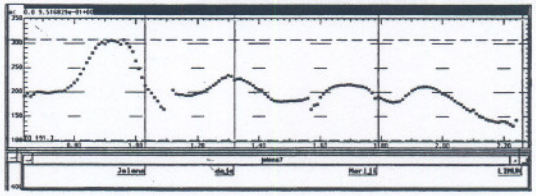


Figure 44: *Jelena daje Mariji limun* 'Jelena is giving Mary a LEMON.'

Looking at the peaks, we may notice that the prosodic focus on *Jelena*, in Figure 41, didn't raise the peak of this word but instead has reduced the peaks of the subsequent constituents. Prosodic focus on the verb seems to have lowered the H of the initial word and also reduced the peaks on the subsequent words. In Figures 43 and 44 , the peaks of the focused constituent seem to be higher than in the broad focus utterance, Figure 40. We can also look at what happens to the final constituent. There seem to be three types of realization of this word: in a broad focus utterance, Figure 40, after a prosodic focus, Figures 41, 42, and 43, and being prosodically focused itself, Figure 44. A broad focus utterance gives the final constituent a reduced pitch range. The constituents following prosodically focused constituent manifest a much flatter pitch line. In other words, narrow focus affects post-focal constituents via pitch range reduction. The pitch range manipulation can be represented in the following way:

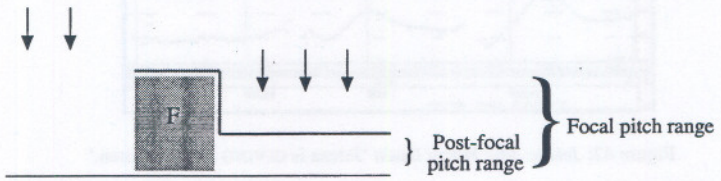


Figure 45: Schematic representation of pre-focal and post-focal pitch range reduction.

Prosodic focus affects the final constituent by widening the pitch range for this constituent, which is the reverse of what is observed for this position when it is not prosodically focused. This expansion of the pitch range for the final constituent allows the manifestation of the lexical accent with no reduction. This is another piece of evidence that the final position does not neutralize the accents (see section 4.2.2).

The next figure shows a familiar utterance from section 4.2.3 with the narrow focus on *žena* 'woman'. Being a longer utterance, the effect of prosodic focus is more obvious in the pitch track of this utterance than in a shorter utterance.

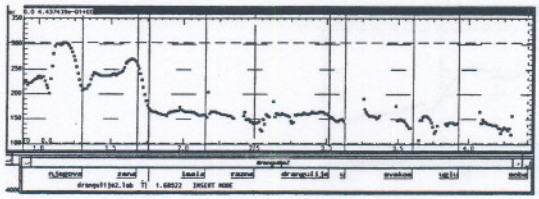


Figure 46: Another illustration of the pitch range reduction after the prosodic focus: *Njegova žena je imala razne drangulije u svakom uglu sobe* 'His WIFE had all sorts of junk in every corner of the room.'

The same effect of prosodic focus is reported for Mandarin Chinese (Jin 1996) and for Hindi (Harnsberger & Judge 1996). Jin shows that post-focused constituents (post-stressed syllables in his terminology) are affected by a significant pitch range reduction, whereas pre-focused constituents are not. According to Harnsberger & Judge (1996), Hindi also signals prosodic focus by drastically reducing the pitch range of the post-focal constituents, a phenomenon which they call register compression.

How can we account for the pitch range effect due to prosodic focus? I propose that focus is signaled by a phrase accent. The L- tone of the phrase accent is realized at the right edge of the word which is focused. That is, the phrase accent is realized earlier than the right edge of its phrase and spans over the string in the post-focal domain, which in turn lowers the pitch range for those constituents.

4.2.5 Morphologically Marked Questions

In this section, I look at the intonation of three types of questions: two types of yes-no questions, both of which employ the question particle *li*, and standard *wh*-questions.

The point of this section is to show that there are no prosodic differences between declaratives (that we have looked at thus far) and morphologically marked questions. That is, there is no special intonation necessary if the interrogative mood is morphologically specified. I look at *wh*-question, and *yes-no* questions.

Unless they are echo-questions, *wh*-questions obligatorily have the *wh*-word at the beginning of the sentence. In syntactic terms, *wh*-movement is obligatory. Grammatical status of the *wh*-constituent, argument vs. adjunct, does not affect the prosody of questions. Since *wh*-words are clause initial, their prosodic pattern is of the sentence initial position, discussed in 4.2.1, as the following pitch tracks show:

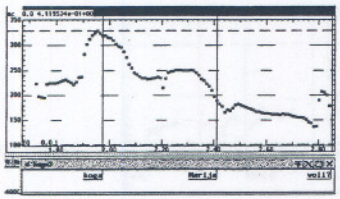


Figure 47: *Koga Marija voli?* 'Who does Mary love?'

Wh-words can also be focused, in which case the prosodic focus effects are the same as in declaratives; the post-focal constituents are in a drastically reduced pitch range. Compare the declarative prosodic focus, Figure 46, and the prosodic foucs on the *wh*-word found in the following figure.

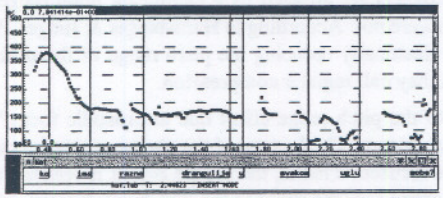


Figure 48: *Ko ima razne drangulije u svakom uglu sobe?* 'Who has all sorts of junk in every corner of the room?'

Yes-no questions are formed in several ways. The standard way is to start the question with *da li* (*Da li Marija voli Milana?* 'Does Mary love Milan?'); *li* is a question particle, and *da* is a complementizer 'that'. Another way is to start the question with *je li* (*Je li Marija voli Milana?* 'Does Mary love Milan?'), *je* is the short (clitic) form of the 3p.sg.pres. of the verb 'to be'. Clitics are by definition unaccented forms; however, when *je* precedes the question particle, it bears a short-falling accent.[9] And finally, the third way is to attach *li* to the tensed verb or some other constituent that is being questioned. For the purpose of illustration, here are some examples of the third strategy:

(5) a. Milan li je otišao?
Milan.NOM li AUX left
'Was it Milan that left?'

[9]That is, in order to support a clitic it must be prosodically "promoted" to an accented form.

b. Ode li Milan?
left.AOR li Milan
'Did Milan leave?'

c. Kući li je Milan otišao ?
home li AUX Milan left
'Was it home that Milan left?'

A pitch track of a standard *da li* question is no different from a simple declarative utterance, as the following figure shows:

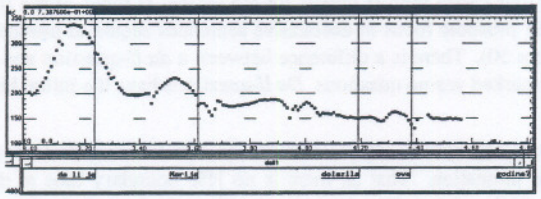


Figure 49: *Da li je Marija dolazila ove godine?* 'Did Mary come this year?'

However, the second type of *yes-no* questions, those with *je li*, seem to favor some additional focal prominence, most particularly on the verb, as also noted by Lehiste and Ivić 1977. Prominence on the verb is also found in Russian morphologically unmarked questions (Ladd 1996). For example:

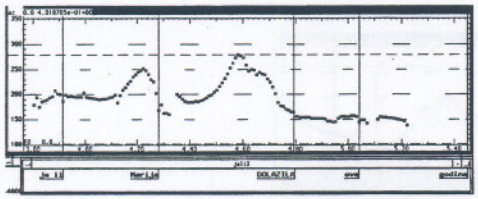


Figure 50: *Je li Marija dolazila ove godine?* 'Did Mary COME this year?'

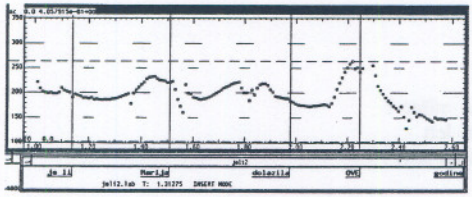


Figure 51: *Je li Marija dolazila ove godine?* 'Did Mary come THIS year?'

Yes-no questions create high H targets for the lexical H tones of the focused constituents, higher than prosodic focus in declarative sentences seems to produce (compare Figure 42 with Figure 50). There is a difference between a *da li*-question and other types of morphologically marked yes-no questions. *Da li*-questions have the initial high rise, just like declaratives.

To summarize, common to all questions is the fact that the final constituents do not exemplify a rise intonation. That is, there is no H% boundary tone at the end of a morphologically marked question. However, as we will see in the next section on question tags and the section on prompting intonation, the H% tone can mark utterances as questions when they are not morphologically marked (just as in English).

4.2.6 Question Tags

Another way to ask a question, employing morphology, is to use a question tag *zar ne?* or *jel' da?* 'isn't it the case?'. The basic contour of these types of questions involves a rising intonation at the end. I use these utterances as evidence for a H% boundary tone of an intonational phrase.

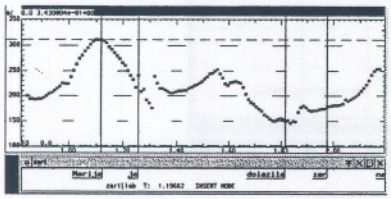


Figure 52: *Marija je dolazila, zar ne?* 'Mary came, didn't she?'

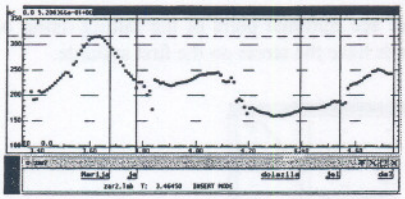


Figure 53: *Marija je dolazila, jel' da?* 'Mary came, didn't she?'

On the basis of contrast between the boundary tones in question tags and declarative utterances, I propose two different right edge intonational phrase boundary tones: L% and H%. Together with the phonological word left edge boundary tone, these tones are some of the markers of prosodic structure.

In the next section I introduce three new markers of prosodic structure: a L- phrase accent and H% boundary tone found in prompting intonation, a %H word boundary tone, found in double focus constructions, and H- phrase accent found in vocative chant. I discuss these contexts in a separate section because their tonal properties interact with lexical tonal specification that leads to loss of lexical information.

5 Loss of Lexical Information

5.1 Prompting Intonation

Prompting intonation can be characterized as the intonation pattern used for elicitation of information about some constituent. For example, it could be the intonation contour on *Marija?!* which can then have the meaning of: 'What about Mary? Tell me something about her.' This intonation pattern can also be used for signaling a yes-no question, or for signaling surprise. L&I have studied this intonation pattern as a question intonation for morphologically unmarked yes-no questions. They name it 'a reverse pattern'. I will continue to call it prompting intonation in accordance with the terminology used by I&Z. As we will see, this intonation pattern seems to neutralize the lexical accents' patterns, the claim also made by L&I:190. Prompting intonation then is an intonational morpheme that seems to overwrite the phonemic distinctions made by the lexical accents.

In a constituent under the prompting intonation there is a steep rise immediately after the stressed syllable. One hypothesis would be that this rise could be represented as a H% boundary tone. However, I will argue that prompting intonation is not just a simple H% boundary tone, but that it is a sequence of L- phrase accent followed by a H% boundary tone. The reason for this will become clear when we look at the pitch tracks of this contour.

In the following figures we see minimal pairs of the falling/rising opposition in prompting intonation. All four words have the stress on the first syllable.

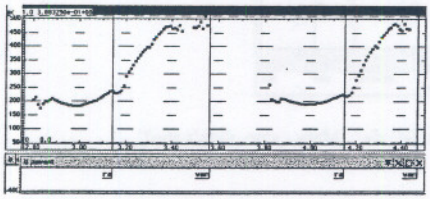


Figure 54: Long Falling/Rising accents: the minimal pair *râvan* 'plain' (falling accent) and it rávan 'flat' (rising accent) in prompting intonation.

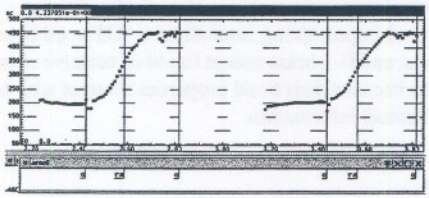


Figure 55: Short Falling/Rising accents: *ȍrao* 'he plowed' (falling accent) and *òrao* 'eagle' (rising accent) in prompting intonation.

The above figures show us that it is very hard to see any distinction among the minimal pairs in falling/rising opposition, as we saw it in declarative utterances. That is, the F_0 of the stressed syllable of the falling accents seems to be very similar (although there are some very small differences) to the stressed syllable of the rising accents in this intonational pattern even though according to their lexical specification we would expect them to be different, as they are in the declarative intonation pattern. Thus, this intonation pattern is a candidate for accent neutralization environment.

According to my data, and also according to L&I's analysis, all the accents seem to be neutralized under the prompting intonation in terms of their F_0 values.

We can see that prompting intonation affects the portion starting at the stressed syllable by looking at words with late accent placement, as illustrated in the pattern on the, by now familiar, word *omalovažavanje*, in Figure 56.

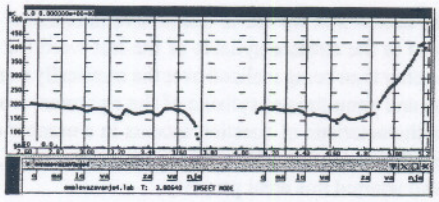


Figure 56: Long-rising accent: *omalovažàvanje* 'humiliation' in citation form and prompting intonation.

The fact that preaccentual syllables are not affected by the prompting intonation allows words with late accents to be more easily distinguished, since only rising accents occur on non-initial syllables.

Monosyllabic words (which can only bear a falling accent) also show a pattern that is hard to account for if we assume that the lexical information is preserved under this intonation pattern, since there is no post-stressed syllable. The following pitch-tracks show the long-falling and the short-falling accent in a prompting intonation of a monosyllabic word.

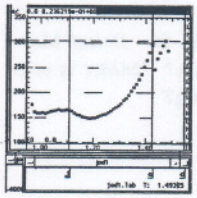


Figure 57: Long-falling accent (*jôd* 'iodine').

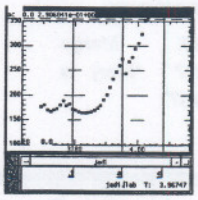


Figure 58: Short-falling accent (*jȁd* 'grief').

As Figures 57–58 show, the prompting intonation can also be realized on a single syllable. The basic pattern of this intonation type is preserved; the super-high target is realized in the second half of the syllable, even in the word under the short-falling accent, which is monomoraic, as discussed in section 3.3.4. These examples provide evidence that tones associated to structurally higher units can overwrite the tonal specification from lower levels.

To account for this intonation pattern I propose a sequence of L- phrase accent followed by a H% boundary tone. There is a difference however in the alignment of this

phrase accent and the L- phrase accent that we see in declarative utterances. The L tone of this phrasal accent is anchored to the stressed syllable of the last word (or the focused word – see Figures 62 and 63), rather than being realized over the metrically non-prominent ultimate (or sometimes ultimate and penultimate) syllable of the rightmost word. Grice et al. (in press) show that this is a characteristic of question accents in a number of unrelated Eastern European languages and their varieties, such as Hungarian, Romanian, and Greek. Serbo-Croatian has evidently also aquired this areal property.

Prompting intonation is also a prosodic focus marker, albeit with a question/surprise semantics rather than emphasis alone. Indicative sentences can be given interrogative mood with this intonation pattern. Figures 59 and 60 are examples of morphologically unmarked yes-no questions under the prompting intonation. Reversing the word order in the question produces a different focus, as indicated by the translation.

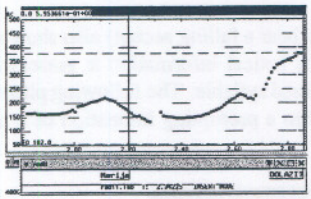


Figure 59: *Marija* DOLAZI*?* 'Mary is COMING?'

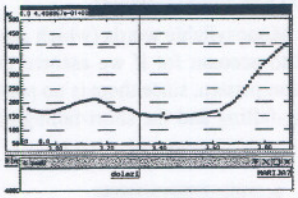


Figure 60: *Dolazi* MARIJA*?* 'MARY is coming?'

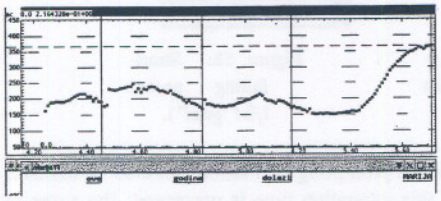


Figure 61: *Ove godine dolazi Marija?* 'MARY is coming this year?'

These examples are interesting because of the interaction between prompting intonation and focus of the question. The focus of the question is the word which bears the phrase accent, i.e. the edge constituent in the above examples. To a limited extent, it is

possible to extend this edge. The following pitch tracks illustrate this point with both rising and falling accents.

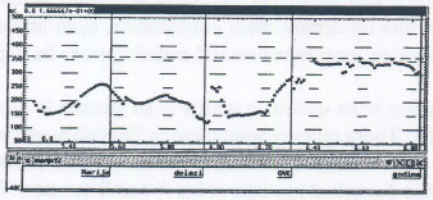


Figure 62: *Marija dolazi òve gòdine?* 'Mary is coming THIS year?'

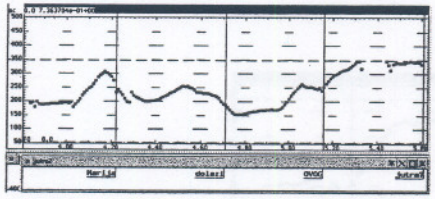


Figure 63: *Marija dolazi òvog jùtra?* 'Mary is coming THIS morning?'

As we can see in Figures 62–63, the focused constituent is under the prompting intonation, and the constituent after it is in a highly raised and compressed pitch range. Falling/rising distinction also seems to be lost in this position. The length of the stretch following the H boundary tone seems to be limited to relatively short strings. As is even more clear in these examples, the L tone is anchored to the stressed syllable of the focused constituent, whereas the H tone is always at the edge. In other words, the two tones are timed differently.

As Ladd (1996) and Grice et al. (in press) show, in Hungarian, Romanian, and Greek questions are marked by the sequence L* H L, where the L* targets the stressed syllable of the focused word, and the HL sequence follows. In Hungarian the HL sequence targets the last two syllables of the phrase, whereas in Greek and Romanian the H tone of this sequence will target a stressed syllable if there are any. Thus, Serbo-Croatian prompting intonation differs from the one in the surrounding languages in the fact that Serbo-Croatian the tonal sequence is bitonal rather than tritonal, as it is in these langauges. Serbo-Croatian does not have the final L boundary tone found in these languages.

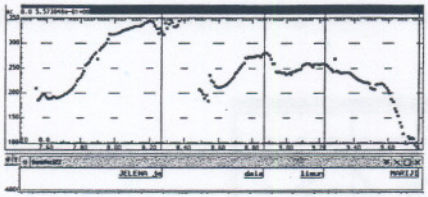


Figure 65: *Jelena je dala limun Màriji.* 'Jelena gave a lemon to Mary.' This utterance was an answer to the question *Ko je kome dao limun?* 'Who gave a lemon to whom?'

When the independent focus is not followed by the dependent focus, there is usually a break between the two phrases. In the above utterance, Figure 65, the dependent focus was placed at the end of the utterance. We can see that even in the final position, which as we have seen is low in broad focus utterances and especially low in non-focused utterances, there are two pieces of evidence for %H boundary tone: (a) there is no dip in the pitch contour signalling the L word boundary and, (b) signaling the finality of the phrase requires a much steeper fall, since the %H word boundary tone has raised the pitch range for the final constituent.

The %H word boundary tone affects the shape of the rising accents of the word to which it is attached, but the falling/rising opposition of accents is still distinguished, as shown by the following pitch track, which has a falling accent on the last word, as opposed to a rising accent in the previous utterance.

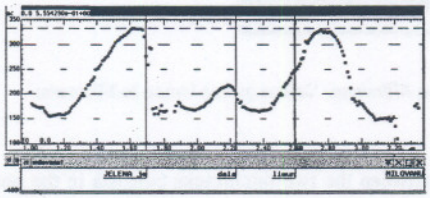


Figure 66: *Jelena je dala lȉmun Mȉlovanu.* 'Jelena gave a lemon to Milovanu.' This utterance was an answer to the question *Ko je kome dao limun?* 'Who gave a lemon to whom?'

However, the lexical accents preceding the %H tone seem to be affected. We can compare the F_0 shape of *lȉmun* 'lemon' in the preceding figure with the F_0 shape of *rávan* 'flat one'

in Figure 67. The two words are similar enough for comparison, but differ in falling/rising opposition.

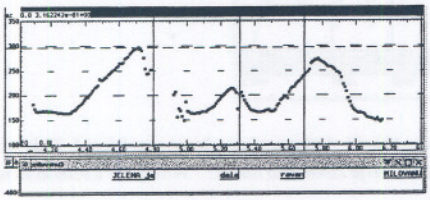


Figure 67: *Jelena je dala râvan Mìlovanu.* 'Jelena gave a/the flat one to Milovanu.' This utterance was an answer to the question *Ko je kome dao ravan?* 'Who gave a/the flat one to whom?'

Compare Figure 66 to Figure 67, which is a broad focus utterance of the same sentence.

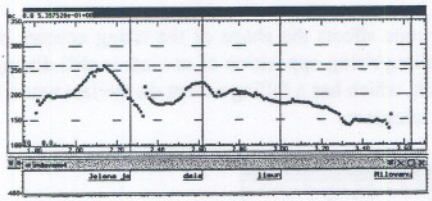


Figure 68: *Jelena je dala lìmun Mìlovanu.* 'Jelena gave a lemon to Milovanu.'

On this analysis the difference between the two pitchtracks consists in the phrasing, one intonational phrase in Figure 68 vs. two intonational phrases in Figure 66. In addition, the tonal strings are also different. In Figure 68, all lexical tones are preserved. In Figure 66, the lexical tones of the word 'lemon' are affected by the %H word boundary tone of the following focal constituent, as the comparison of the two figures clearly shows.

Thus, I conclude this section by noting that the %H word boundary tone affects the lexical information of preceding constituent. I now turn to the last section in which I discuss the vocative chant intonation.

5.3 Vocative Chant

Vocative chant is another intonational contour which seems to affect the lexical information to a great degree. The melody that characterizes vocative chant is similar but not identical to the English vocative chant. I&Z observe that 'the vocative chant has a basic (Low)-High-Mid melody', where the Low is present only in words with three syllables or more and with the stress on non-initial syllable.

According to my findings, the initial L tone is present in all cases, which in this system is accounted by the %L word boundary tone. The rest of the shape of the F_0 contour shows a rise towards a H target and a continuation with a slight drop in pitch. This basic pattern shows up on all words regardless of their metrical structure. That is, what is common to all words under vocative chant intonation pattern is the H tone on the penultimate syllable and a lower tone on the final syllable. The fact that the two tones go together and target the last two syllables of the word argues in favor of an analysis which treats this pattern as a property of the phrase edge, i.e., a boundary tone.

The vocative chant melody can be seen in the following pitch tracks of trisyllabic words with the stress on the first syllable:[10]

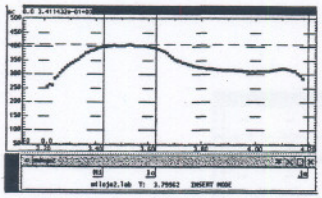


Figure 69: *Mìloje!*

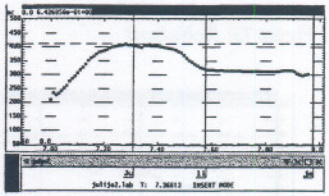


Figure 70: *Jûlije!*

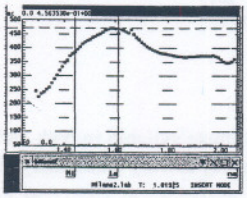


Figure 71: *Mìlane!*

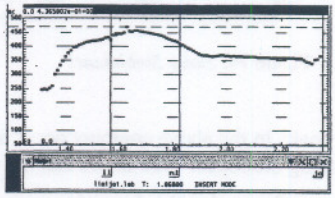


Figure 72: *línijo!*

[10]At this time I don't have an example of an all sonorant trisyllabic proper name under the long-rising accent, so I have used a common noun for illustration purposes.

It seems quite obvious that vocative chant is affecting the lexical specification for the tonal information. However, it is not entirely clear that the lexical information is completely lost. There seems to be at least one difference between falling and rising accents. Rising accents have a slightly higher target for the H tone in the above examples (compare Figures 69 and 70 with Figures 71 and 72). Whether the lexical H tone which correlates with the second syllable is boosting the boundary H tone is an open question and would require a detailed study.

To show that vocative chant is a boundary effect, we can look at examples of longer words. In the following two figures I show a calling contour on a word *Slobòdane!* and its possible extended variant, *Slobo-Slobòdane!* .

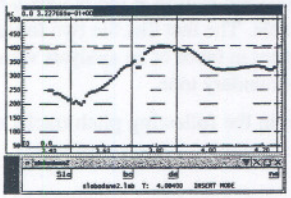


Figure 73: *Slobòdane!*

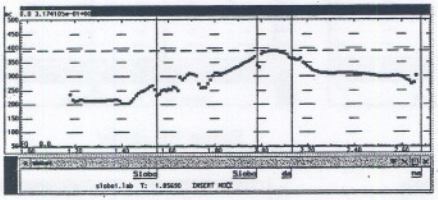


Figure 74: *Slobo-Slobòdane!*

F_0 peaks in the above contours occur on the penultimate syllable in both variants of summoning Slobodan. This shows that the vocative melody is truly a boundary effect.

To see that this bitonal boundary tone is targeting the last two syllables of a word regardless of the position of the stressed syllable, we can look at a stress initial word with more than three syllables. Consider the following pitch track.

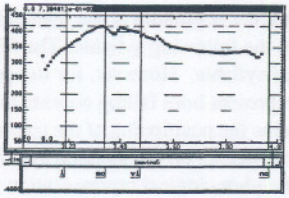


Figure 75: *ìmovino!*

Figure 75 shows again that the H tone is associated with the third, penultimate syllable, even though the stress is on the first syllable. Because the stress syllable does not play a role anchoring the melody, I conclude that vocative chant can be analyzed as a H- phrase accent followed by a L% boundary tone. This melody then differs from the prompting intonation where we saw that L- phrase accent is a special kind of boundary tone because it is timed to the stressed syllable.

There is one more thing to mention regarding vocative chant. The L tone of this boundary tone is not as low as the single L% boundary tone that we see in declarative utterances. It may seem reasonable then to question this characterization of this tone as a L. Presumably, another possible analysis for this contour would be to say that we have H- phrase accent followed by a downstepped !H% boundary tone. So far, we have no other evidence for a downstepped boundary tone. At this point then it seems unjustified to introduce a new target just for this melody. However, should such evidence arise, a reanalysis may be appropriate.

To sum up, in this section we have seen three different types of structural markers: L- phrase accent followed by a H% boundary tone (prompting intonation); a %H word boundary tone of double focus constructions, and a H- phrase accent followed by a L% boundary tone (vocative chant). It seems that the unifying property of these edge tones is that structural H tones affect lexical information so that lexical pitch accents end up neutralized. This is still a tentive conclusion until more data become available.

6 Conclusion

The surface tones of Serbo-Croatian accents can be described in terms of alternations between H and L tones, with some tones assigned in the lexicon, others assigned at the level of prosodic phrasing and yet others functioning to integrate pragmatic coherence of the discourse. I have argued that falling accents have a HL melody whereas rising accents have LH melody. The difference between the short and long falling is in the timing of the fall, which is the function of the length of the tone bearing unit (mora) to which the

H tone is anchored to. Long falling accents have the beginning of the realization of the fall on the stressed syllable itself and continuing on the following syllable. The fall of the short falling accent is delayed until the post-stressed syllable. However, for the purpose of the phonological representation it is sufficient to represent both falling accents as H*+L, since the duration of the stressed syllable will determine the positioning of the trailing tone. Both rising accents show the LH melody where the L occurs on the stressed syllable and the H on the post-stressed syllable. Consequently, the phonological representation for the rising accents is L*+H. The timing of the trailing H tone of the rising accents also does not need to be stipulated. Anchoring the L tone to the (last) mora of the accented syllable produces the desired effect of having the H tone on the post-stressed syllable and yet gives us enough flexibility, as with the falling accents, to accomodate variations in production.

The short/long distinction between the rising accents also seems to be accompanied by a difference in vowel quality. Thus it might be necessary to include a study of vowel quality together with the lexical accentual properties. Therefore, the proposal offered here for the description of the accents only in terms of a two way distinction, (falling vs. rising) may be necessary and sufficient.

A broad focus declarative utterance allows all lexical tones to be realized. Phonological words are clearly separated by %L, a word boundary tone. In addition, each subsequent phonological word is down-stepped from the previous one. The sentence initial constituent, regardless of its syntactic function, is set off from the rest of the constituents by having the highest target for the realization of the lexical H. The sentence final constituent is conversely in the lowest pitch range. Nevertheless, the realization of the lexical accents is still present. The falling accents in this position show a steady fall in the pitch, whereas the rising accents maintain the same pitch level in the post-stressed syllable, thus marking the two accents differently. Focusing the final constituent in a sentence allows all lexical tones to be fully realized, providing additional evidence that the phonological representation is not lost in final position.

In prosodic narrow focus utterances, the constituents following the focused constituent are in a markedly reduced pitch range relative to an utterance without the prosodic focus. In a paradigmatic contrast with broad focus utterances, prosodic focus slightly expands the pitch range of the focused constituent, creating a higher target for the lexical H, and compresses the pitch range surrounding the focused constituent, most drastically the following ones. That is, focal prominence involves not so much making the focal peak higher as it does make non-focal peaks lower. I have proposed that post-focal pitch range reduction is a consequence of the early realization of the L- phrase accent at the right edge of the focused word. The double focus construction provides evidence for a %H word boundary tone, which is used to signal the dependent variable constituent in this construction. This boundary tone differs from pitch range expansion in narrow focus constructions in the following way: pitch range expansions provides a wider tonal space for all tonal

targets, %H word boundary tone raises the tonal target of the left edge of the word thereby creating a pull for the preceding and subsequent L tones.

Morphologically marked questions do not have a H% boundary tone, whereas non-marked questions can be signaled either by a question-tag wich always has a H% boundary tone, or by prompting intonation on the focal constituent. Prompting intonation and vocative chant is a result of the combination of a phrase accent followed by a boundary tone, L- H% and H- L% respectively. Both of these intonational contours seem to affect lexical pitch accents.

To sum up: in this paper I have argued for three levels of prosodic phrasing in Serbo-Croatian, a phonological word, an intermediate phrase and an intonational phrase. The two prosodic units are either associated with certain tonal markings, such as edge tones, or function as a domain of a rule application. A phonological word has a delimitative marker, an initial wordy boundary tone, which can be either %L or %H, and a culminative marker, which can be any of the four pitch accents. The intermediate phrase is marked by phrasal accents (L- and H-) and the intonational phrase has two types of boundary tones (L% or H%).

Since Serbo-Croatian is both a stress language and a pitch accent language, it provides an example of a very different type of language than the ones that have been studied in depth so far from an intonational point of view, such as English, Japanese and others. In particular, I hope to have shown that one of the main points of interest in study of Serbo-Croatian intonation is the interaction of lexical tonal specifications with the tonal markings of intonational phrasing.

REFERENCES

Beckman, M. and J. Pierrehumbert (1986). 'Intonational Structure in Japanese and English'. Phonology Yearbook 3:255-309.

Brown, G., K. L. Currie, & J. Kennworthy (1980). *Questions of Intonation*. London: Croom Helm.

Browne, E. W. (1967). 'On the Problem of Enclitic Placement in Serbo-Croatian'. MIT ms.

Browne, E. W. and J. D. McCawley (1965). 'Srpskohrvatski akcenat'. Zbornik za filologiju i lingvistiku 8:147-151.

Bruce, Gösta (1977). *Swedish word accents in sentence perspective*. Lund: Gleerup.

Bruce, Gösta (1990). 'Alignment and composition of tonal accents: comments on Silverman and Pierrehumbert's paper', in Beckman M. and J. Kingston (1990) *Between the grammar and the physics of speech* Papers on Laboratory Phonology: 1. Cambridge: Cambridge University Press.

Clements, G. N. and K. C. Ford (1979). 'Kikuyu Tone Shift', LI 10.2.

Clements, G. N. and K. C. Ford (1981). 'On the Phonological Status of Downstep in Kikuyu', in D. L. Goyvaerts ed. (1981). *Phonology in the 1980's*. Ghent/Belgium: E. Story-Scientia.

Godjevac, S. (1999). 'Declarative Utterance-Final Position in Serbo-Croatian' in Osamu Fujimura, Brian Joseph and Bogumil Palek (eds.) Proceedings of LP'98. Prague: The Karolinum Press. 63–76.

Grice, M., R. D. Ladd, and A. Arvaniti (in press) "On the place of phrase accents in intonational phonology". Phonology. Cambridge: Cambridge University Press.

Grønnum, N. (1985). 'Intonation and text in Standard Danish'. JASA 77(3):1205-1216.

Grønnum, N. (1992). *The groundworks of Danish intonation: an introduction*. Copenhagen: University of Copenhagen/Museum Tusculanum Press.

Gvozdanović (1980). *Tone and Accent in Standard Serbo-Croatian*. Verlag Der Österreichischen Akademie Der Wissenschaften Wien.

Halle, M. (1971). 'Remarks on Slavic Accentology'. LI I.2

Harnsberger, J. D. and J. Judge (1996). 'Pitch Range and Focus in Hindi', paper given at 131st meeting of the Acoustical Society of America in Indianapolis.

Hirschberg, J. & J. Pierrehumbert (1986). 'The Intonational Structuring of Discourse'. Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics.

Inkelas, S. and D. Zec (1988). 'Serbo-Croatian Pitch Accent: The Interaction of Tone, Stress, and Intonation'. Language 64.2:227-248

Inkelas, S. and D. Zec (1990), *The Phonology-Syntax Connection*, CSLI, The University of Chicago Press, Chicago.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, Mass. MIT Press.

Jin, Shunde (1996). An Acoustic Study of Sentence Stress in Mandarin Chinese. PhD dissertation, OSU.

Jun, Sun-Ah & Mira Oh (1996). 'A Prosodic Analysis of Three Types of Wh-Phrases in Korean'. Language and Speech 39(1):37-61.

Kariya, C. S. (1983). *The Acquisition of Accent in Serbo-Croatian: A Case Study*. PhD Thesis, Harvard University.

Kostić, Dj. (1983). *Rečenička melodija u Srpskohrvatskom jeziku*. Beograd: Rad.

Kubozono, H. (1992). 'Modeling syntactic effects on downstep in Japanese' in Docherty, G. J. & R. D. Ladd (eds.) *Gesture, segment, prosody - (Papers in laboratory phonology v.2)*. Cambridge: Cambridge University Press.

Ladd, R. (1984). 'Declination: a review and some hypotheses' Phonology Yearbook 1:53-74.

Ladd, R. (1988). 'Declination 'reset' and the hierarchical organization of utterances'. JASA 84:530-44.

Ladd, R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Lehiste, I. (1975). 'The phonetic structure of paragraphs', in A. Cohen & S.G. Nooteboom (eds.) *Structure and Process in Speech Perception*. Berlin: Springer-Verlag, pp.195-206.

Lehiste, I. and P. Ivić (1963). *Accent in Serbocroatian: An experimental Study*. Michigan Slavic Materials 4. Ann Arbor: University of Michigan, Dept. of Slavic Languages and Literatures.

Lehiste, I. and P. Ivić (1977). 'Interrelationship between Word Tone and Sentence Intonation in Serbocroatian' in Donna Jo Napoli ed. *Elements of Tone, Stress, and Intonation*. Georgetown University Press, Washington, D.C.

Lehiste, I. and P. Ivić (1986). *Word and Sentence Prosody in Serbocroatian*. Cambridge, Massachusetts: The MIT Press.

Liberman, M. & J. Pierrehumbert (1984). 'Intonational Invariance under Changes in Pitch Range and Length' in M. Aronoff & R. T. Oerhle *Studies in Phonology. Presented to Morris Halle by his teacher and students* Cambridge, Massachusetts: The MIT Press.

Maekawa, K. (1991). 'Perception of intonational characteristics of WH and non-WH questions in Japanese'. Proceedings of the XIIthe International Congress of Phonetic Sciences, 4/5:202-205.

Nespor, M. & I. Vogel. (1986). *Prosodic Phonology*. Dordrecht: Foris.

Nikolić, B. (1970). *Osnovi Mladje Novoštokavske Akcentuacije*. Institut za Srpskohrvatski Jezik, Beograd.

Pierrehumbert, J. (1980). *The phonetics and phonology of English intonation*. PhD dissertation, MIT.

Pierrehumbert, J. & M. Beckman (1988). *Japanese Tone Structure*. Cambridge, Massachusetts: The MIT Press.

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

Selkirk, E. (1986). "On derived domains in sentence phonology". Phonology Yearbook 3: 371-405.

Stevanović, M. (1989). *Savremeni Srpskohrvatski Jezik*. Beograd: Naučna Knjiga.

Venditti, J. (1994). 'The influence of syntax on prosodic structure in Japanese'. OSU Working Papers in Linguistics 44:191-223.

SOUND CHANGE ACROSS SPEECH ISLANDS: THE DIPHTHONG /aɪ/ IN TWO MIDWESTERN PENNSYLVANIA GERMAN COMMUNITIES*

Steve Hartman Keiser

Abstract

This paper analyzes the variable production of the Pennsylvania German diphthong /aɪ/ in two Pennsylvania German speech islands in Iowa and Ohio. The data show that younger speakers regularly monophthongize /aɪ/, yielding [ɛː] or even (in Ohio only) [eː], and perceptual studies show that the latter form merges with the vowel space of the phoneme /eː/. This sound change is shown to be an example of language drift (i.e., internally motivated), though its spread across distant speech islands is suggestive of significant ongoing patterns of interaction between these speech islands.

0 Introduction

This paper presents evidence for a sound change in progress in the vowel system of Midwestern Pennsylvania German (PG): the monophthongization and fronting/raising

* Thanks to Brian Joseph, Don Winford, and Rich Janda for their comments and suggestions. Also thanks to Keith Johnson for advice on perception experiments, to Anna Grotans for references on German etymology, Mary Beckman, Matt Makashay, Christian Uffman, and many others in the Linguistics Department at OSU. Finally a big thank you to the Matthew Schrock extended family for hours of conversation and hospitality, and to the dozens of coworkers who patiently taught me and became friends in addition to serving as an invaluable source of data during our days together.

of the diphthong /aɪ/. Variation in the phonetic production of this phoneme is socially significant. The use of the (older) variant [aɪ] in a word such as [daɪtʃ], 'German,' is described as non-native or typical of a second-language learner. In addition, a subset of speakers in one Midwestern community produce variants of /aɪ/ that overlap the vowel space of the PG phoneme /eː/, and preliminary perceptual testing indicates that phonemic merger is underway. The fact that this sound change has spread across geographically distant communities poses questions for processes of dialect contact across speech islands.

I begin with a review of the previous research on this phenomenon. In the second section I provide a brief synchronic description of the PG vowel space and delineate the word set containing /aɪ/ in the Kalona and Holmes County dialects. I also present data on the production of /aɪ/ from earlier time periods to establish the diachronic basis for the sound change. I introduce my synchronic data in the third section, including a description of the selection of variants and an investigation of the linguistic and social conditioning of these variants. I then investigate a possible phonemic merger underway and test its salience. The fifth section I devote to discussions of various accounts for the origin and spread of the sound change. Finally, I comment on some implications of these data for the study of the spread of sound change between geographically noncontiguous communities.

1 Previous Research

To date, only two researchers have mentioned the vowel system developments in question here. In Schlabach's 1980 thesis on the phonology of Holmes County PG, he comments: "...some speakers (as I have observed) regularly substitute the long vowel /æː/ for the diphthong /aɪ/ in all words in OPG" (39). He goes on to note the following examples, including some minimal pairs distinguished by nasalized vowels (39, 43):

(1) /haɪt/ ~ /hæːt/ 'today' (39)
/daɪ/ ~ /dæː/ 'your' (sg.) (43)
/saɪ/ ~ /sæː/ 'pigs' (67)
/saɪ/ ~ /sæ̃ː/ 'his' (67)
/naɪ/ ~ /næː/ 'new' (67)
/naɪ/ ~ /næ̃ː/ 'in' (67)
/naɪn/ ~ /næːn/ 'nine' (67)

Schlabach appears to restrict this variation to speakers of the Madison County dialect (5, 42). However, some of Schlabach's data suggest that this variation may be more widespread than that. These data describe a monophthongal production [æː] or [eː] for words which other Ohio PG sources describe as [aɪ].

(2)	Schlabach data	Data in *Es Nei Teshtament* (ENT)
	/væːl/ 'because' (35)	/vaɪl/
	/faːlheːt/ 'laziness' (49)	/vaɪshaɪt/ 'wisdom' (affix -/haɪt/)

Louden (1997, 81) is the first to give an account of this change in dialects outside of Ohio. He describes the monophthongization of /aɪ/ to /ɛː/ as a system-internal balancing of front and back long vowels and notes that this change in progress is farther advanced in Midwestern[1] PG than in Lancaster County, Pennsylvania (see (3) below).

(3)	Lancaster rule:	Monophthongize only before liquids	
		/aɪ/ > [ɛː]/___[ɹ,l]	e.g., [miɐ hɛːɹ] 'we marry'
		[aɪ] elsewhere	e.g., [daɪtʃ] 'German'
	Midwestern rule:	Retain diphthong only before unstressed central vowels.	
		/aɪ/ > [aɪ]/____ [ə, ɐ]	e.g., [miɐ haɪərə] 'we marry'
		[ɛː] elsewhere	e.g., [dɛːtʃ] 'German'

Louden's account rests crucially on a characterization of the PG vocalic system with reference to quantitative (long/short) rather than qualitative (tense/lax) differences and also on the notion of symmetry as an organizing principle for vocalic systems. I will give some consideration to the quantitative vs. qualitative nature of the PG vocalic system in the following section.

2 Synchronic description of PG vowels and the /aɪ/ word class

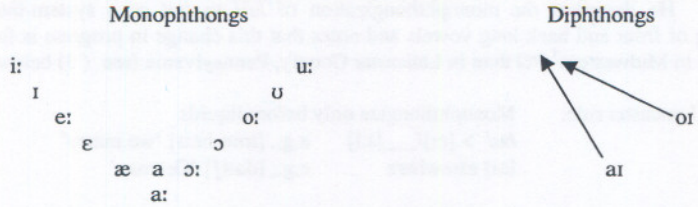
2.1 PG vowels

The following description of six short vowels, six long vowels, and two diphthongs is adapted from descriptions in several sources.[2]

[1] Louden's "Midwestern PG" appears to be a catch-all label for varieties of PG outside Pennsylvania, although he does not say which communities he sampled. The largest Old Order Amish, and hence PG-speaking, communities are in what is commonly considered the "Midwest": in Ohio, Indiana, Illinois, and Iowa. It is not clear, however, if PG varieties spoken in Ontario or Kansas or even central Pennsylvania, for example, are included here. In this paper, I define Midwestern PG as that of Holmes County, Ohio and Kalona, Iowa in opposition to Pennsylvania.

[2] Sources consulted were: Beam (1991, vi), Druckenbrod (1994, 18-19), Frey (1985, 1-2), Van Ness (1994, 422-3), Buffington and Barba (1954, 5-6), Meister Ferré (1994, 19 & 22), Van Ness (1990, 31-9), Schlabach (1980, 30-47).

Figure 1. PG Vowel Space



PG researchers have generally kept with German tradition, describing the PG vowel system as having an opposition based on quantity: a series of long and a series of short vowels. Only Van Ness (1994, 422) suggests that vowel quality is a better descriptor. However, the development of the former diphthong /aʊ/ to present-day /aː/ provides a reasonable argument in favor of a quantitative opposition, at least for low vowels. Currently long /aː/ is in opposition to short /a/ producing minimal pairs such as /has/ 'hate' and /haːs/ 'house' which differ only in length.

2.2 Defining the /aɪ/ word class in PG

The PG diphthong /aɪ/ is generally the reflex of Middle High German (MHG) long, high monophthongs /iː/ and /yː/, e.g., PG /maɪ/ < MHG /miːn/ 'my' and PG /naɪ/ < MHG /nyːwe/ 'new'[3].

I verified the status of the /aɪ/ word class in the lexicon of Holmes County PG by consulting two current texts: the New Testament Bible in PG, *Es Nei Teshtament* (ENT) completed by SIL translators in the mid 1990s, and *Vella Laysa* (VL), a collection of Bible stories written in 1997 by New Order Amish with some initial assistance from SIL translators. . Given that the PG-speakers who served as consultants for ENT were all older men, and that it is considered a sacred text, we can assume that ENT reflects somewhat conservative norms (at least mid-20th century usage or earlier) for the community.[4]

---

[3] These vowels reflect the inventory of the classical period of MHG, defined as 1170-1250 AD by Russ (1982, 60). In fact, diphthongization only affected /iː/, since the round vowel /yː/ unrounded to merge with /iː/ in the Palatinate dialects (the primary input dialects to PG) as early as the end of the 13th century. Diphthongization of /iː/ to /aɪ/was complete before the 16th century (Reed 471). Unlike the dialects upon which Standard New High German is based, the PG source dialects did not collapse reflexes of /iː/, with reflexes of the MHG diphthong /eɪ/. In PG, as in parts of the Palatinate, MHG /eɪ/ yields the monophthong /eː/, e.g., PG /ʃteː/ < MHG /ʃteɪn/ 'stone' (Reed 1972, 472)

[4] One example of this is the use of <au> to represent the diphthong /aʊ/ in spite of current norms of usage which realize this phoneme as monophthongal /aː/, e.g., <haus> for /haːs/ 'house', <naus> for /naːs/ 'out.' A second example is the use of dative morphology in ENT. In conversational speech, dative forms are currently found only in the PG of speakers over the age of 70.

In both ENT and VL, the orthographic symbol for /aɪ/ is <ei>. Several examples are noted in Table 1 below.

Table 1. Example words with <ei> (/aɪ/) listed according to MHG source*

from MHG /iː/	from MHG /yː/	from loss of /r/ or /g/ in /Vri/ and /Vgi/
veisa 'show' 13*	greitz 'cross' 5	deich 'through' 4
shmeisa 'hit' 25	leit 'people' 7	reiyahra 'to rain' 13
zeit 'time' 7	frei 'free' 7	keiyaht 'married' 18
shreives 'writings' 7	eiyah 'your' pl. 7	leisht 'lie (2SG)' 23
	heit 'today' 7	meiyet 'morning' 26
	Deitsh 'German' 7	shteikah 'strong' 44

*numbers indicate page in *Vella Laysa*

I checked these words against the lexical entries in two PG-English dictionaries, Stine 1990 and Beam 1991. All words spelled <ei> in Stine and Beam are also spelled <ei> in the Ohio sources, and the same diphthongal form /aɪ/ is given in all sources[5]. Working from this comparison, it is reasonable to assume that lexical entries with <ei> in Stine and Beam also belong to the /aɪ/ word class in the Holmes County dialect. These dictionaries allow for the easy development of a larger corpus of /aɪ/ words for further analysis.

3 The Data

3.1 Data collection methods and sample size/description.

During fieldwork in Kalona, Iowa (1996) and in Holmes County, Ohio (1998) I conducted one hundred forty standard sociolinguistic interviews which included a translation task. The translation task in Kalona yielded approximately five to seven tokens per speaker. The translation task in Holmes County was longer yielding approximately fifteen to eighteen tokens per speaker. I also recorded casual conversation in a number of settings in homes as a guest and/or co-worker. From these recordings I coded a total of 1187 tokens of words in the /aɪ/ word class from ninety-one speakers.

3.2 Establishing variants of PG /aɪ/ and means of identifying.

In order to develop a scale by which to identify degrees of fronting, raising, and/or monophthongization of /aɪ/, I listened to approximately 50 tokens produced by five different speakers and attempted a narrow transcription which I compared against measurements of F1 and F2 in a spectogram of the utterance.

[5] Approximately twenty words spelled <ei> in ENT and VL are *not* spelled so in Stine and Beam. All but one of these twenty words belong to a set of relatively recent additions to the /aɪ/ word class which are the result of intervocalic weakening and eventual loss of /r/ or /g/ as in <schtarick> 'strong' in Stine and in Beam, written as <shteig> in ENT. Thus, the /aɪ/ word class in Holmes County is larger than the one developed from Stine or Beam, because of the addition of words such as <shteig>.

The salient characteristics for distinguishing vowel quality were height and diphthongal vs. monophthongal status. With respect to measures of tenseness (peripherality in the vowel space), all of the tokens were relatively tense. I employed a four-point scale for vowel height which mirrors the low and front areas of the PG vowel space: /a, æ, ɛ, e/. For diphthongal status I developed a three-point scale which can be further broken down in to two parts: first monophthong vs. diphthong, and second, within the category diphthong, upgliding vs. ingliding, e.g., [æɪ] vs.[æə].

Each token received two ratings: one for height and one for di-/monophthongal quality. The higher the vowel, the more "advanced" the token in terms of change away from a low central nucleus for the diphthong. Both the monophthongs and the ingliding diphthongs can be considered "advanced" tokens in comparison with upgliding diphthongs, though some speakers produce a very salient second ingliding element—in some cases almost a syllabic element—that may represent the most advanced tokens.

Examples are given in Figure 2-Figure 5, below.

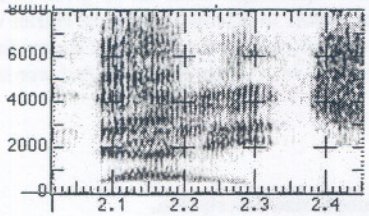


Figure 2. [aɪ] in /daɪtʃ/ 'German', 30 yr old OOA male

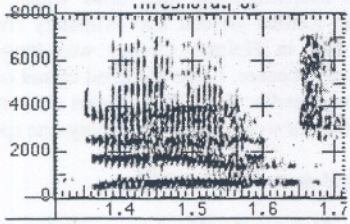


Figure 3. [æ:] in /daɪtʃ/ 'German', 29 yr old OOA male

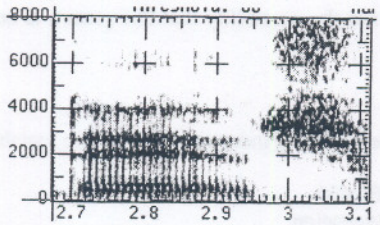


Figure 4. [ɛː] in /daɪtʃ/ 'German', 18 yr old OOA male

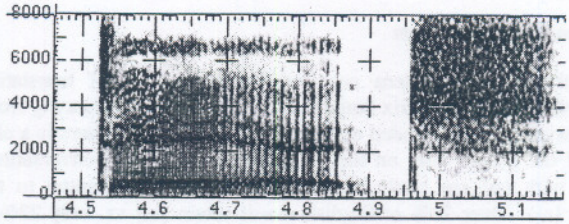


Figure 5. [eə] in /daɪtʃ/ 'German', 32 yr old NOA female

3.3 Variation

The overall distribution of the independent variables vowel height and diphthongal status can be seen in Table 2 below.

Table 2. Diphthongal status vs. vowel height for all data

	vowel height				
	[a]	[æ]	[ɛ]	[e]	TOTAL
monophthong	34	269	348	29	680
upglide	291	36	22	8	357
inglide	5	40	74	31	150
TOTAL	330	345	444	68	1187

Of the twelve possible combinations of the two dependent variables, the most frequently occurring variant is the monophthong [ɛː] (348/1187=29% of total tokens), followed by the conservative diphthongal variant [aɪ] (25%), and the monophthong [æː] (23%). Together these three token types comprise over 75% of the tokens.

So a clear pattern emerges. If a speaker does not produce the canonical [aɪ] token type, then she or he is likely to produce a fronted and perhaps raised monophthong in its

place. The following section explores the possibility that this pattern is conditioned by linguistic variables.

3.3.1 Linguistic variables

Each token was coded for the following independent linguistic variables:

1. language of lexical item: PG or English
2. style: translation task or free conversation
3. lexical item
4. preceding segment
5. following segment

3.3.1.1 Language of lexical item

All of the recorded tokens occurred in the context of translation tasks or conversations with PG as the matrix language. Since PG borrows heavily from American English, each lexical item was coded as either PG or English[6]. There is a clear effect of the language of the lexical item on the vowel quality. In PG conversation, borrowed English words with /aɪ/ are rarely monophthongized and/or fronted to the common variants [æː] or [ɛː]. Over 70% of English words retain [aɪ] vs. only 22% [aɪ] for PG words. This finding suggests that for these bilingual speakers PG phonology and English phonology operate relatively independently of each other.

3.3.1.2 Style

With respect to vowel height, the free conversation data yield slightly more conservative forms than the translation task data. That is, in free conversation, the percentage of [aɪ] tokens increased in both Kalona (from 35% to 53%) and Holmes County (from 20% to 25%). This is perhaps due to the artificial environment of the translation task where borrowed English words were less likely to appear both due to the content of the task and its purpose.

3.3.1.3 Lexical item

This factor is included simply to flag any lexical entries which are unusually progressive or conservative with respect to the sound change. Several words standout as favoring advanced variants, e.g., /gaɪl/ 'horses,' the only lexical item for which a plurality of speakers produced [e]. Given the nature of the corpus, that being that the majority of tokens come from a few high-frequency lexical items (12 words account for

[6] The distinction native vs. non-native vocabulary is very problematic in intense language contact situations such as those in all PG-speaking communities. Here the imperfect criterion used was entry in the dictionary. If a word was listed as a PG entry in Stine and/or Beam, it was labeled a PG word. Thus, some long-term borrowings are considered part of the PG lexicon, e.g., the noun *pie* and the verb *quilt*. Words with inflectional affixes (e.g., plural *–s*) were also included as native PG, while those with derivational affixes (e.g. nominalizing *–ing*, in the gerund *pricing*) were not. For words not listed in Stine or Beam the default classification was English.

approximately 75% of the total tokens) it is difficult here to separate out the possible lexical effects from effects of phonetic environment, e.g., following lateral.

Table 3. Lexical items favoring a particular vowel height variant*

favor [a]	/naɪ/ 'new,' /flaɪxt/ 'maybe,' /ʃaɪɐ/ 'barn,' /kaɪɐt/ 'married'
favor [æ]	/ʃmaɪst/ 'throws (3SG),' /daɪx/ 'through'
favor [ɛ]	/tsaɪt/ 'time,' /daɪtʃ/ 'German,' /haɪt/ 'today,' /daɪ/ 'your,' /saɪ/ 'his,' /draɪ/ 'three,' /glaɪx/ 'like,' /kaixt/ 'obeyed,' /haixə/ 'to obey'
favor [e]	/gail/ 'horses'

*more than five tokens and majority or plurality of tokens produced at one particular vowel height.

The words which most favor the advanced ingliding diphthong variant (inglide occurs in at least 33% of the tokens of the word) are: /tsaɪt/ 'time,' /haɪt/ 'today,' /laɪt/ 'people,' /daɪtʃ/ 'German,' /gail/ 'horses,' /nain/ 'nine.'

3.3.1.4 Preceding and following phonetic conditioning

The segments were coded for preceding and following segmental environments. The nature of the following segment affects the frequency of occurrence both of vowel height and di-/monophthongal quality. In Table 4 and Table 5 a (+) means that there was above-average frequency of the dependent variable in that phonetic environment, a (-) means below-average frequency, and a blank indicates no effect either way. Some strongly disfavoring environments are noted by the label "Ø tkns" which means that no tokens were found in these environments.

Table 4. Effect of following phonetic environment on vowel height

	[ə] or [ɐ]	labial	coronal	palatal	velar	glottal	nasal	morph bndry
[a]	+	+						
[æ]	Ø tkns		+	-		Ø tkns	+	
[ɛ]				+	+			
[e]	Ø tkns	-	+			Ø tkns		-

Table 5. Effect of following phonetic environment on diph-/monophthongal quality

	[ə] or [ɐ]	labial	coronal	palatal	velar	glottal	nasal	morph bndry
monop								+

upglide	+	+	-	-	+	Ø tkns	-
inglide	Ø tkns	Ø tkns	+	+	+	+	-

A following unstressed vowel favors the conservative [aɪ] variant, which supports Louden's (1997) analysis (see section 1). Labials also favor the [aɪ] variant. The most common variants [æː] and [ɛː] are favored by following coronals and palatals respectively. The very advanced form [eə] also is favored by coronals (especially lateral segments) and strongly disfavored by following labial or velar segments. Preceding and following nasal segments favor the [æː] variant and ingliding variants [ɛə] and [æə].

Again, the presence of several high-frequency lexical items in the corpus is cause for caution in interpreting the above findings. The apparently significant effect of phonetic environment might possibly be a lexically-restricted phenomenon.

3.3.1.5 Is the variation of /aɪ/ regular?

The evidence for strictly phonetically conditioned variation is not conclusive since we lack sufficient tokens of particular phonetic environments across different lexical items (particularly preceding /l/ which appears to favor advanced tokens). Still, the data in the preceding section suggest that variation in the production of /aɪ/ is subject to a certain amount of predictable linguistic conditioning, typical of a regular sound change in progress.

3.3.2 Social variables

Each speaker was coded for the following social variables:

1. Community: Kalona, Holmes County, or Pennsylvania (one speaker).
2. Age: a continuous variable which was recoded into four generational cohorts of twenty years each: 0-20, 21-40, 41-60, 60+.
3. Sex: female or male.
4. Denomination: Old Order Amish, New Order Amish, Beachy Amish, Conservative Mennonite, Mennonite.
5. Job: retired, homemaker, factory, office, farmer, student/teacher, small business.
6. Work network: There were three in the Holmes County study. Laborers at the main woodworking factory, office workers at the factory, and installation workers at the factory.
7. Church network: This is basically a geographical measure, since for the most part the Amish go to church with their neighbors. There are 21 of these networks represented, 13 of which are Amish.
8. Family network: There are six families which have three or more members included in the study. There are an additional five which have at least two.
9. Dative usage: Individual's use of tokens of dative morphology in the translation task (part of a previous study) was entered as a continuous variable. This was done in order to test whether conservative usage of a morphological variable (dative case) correlated to conservative usage of a phonological variable (i.e., /aɪ/).

The variables "job," "church network," "family network," and "dative usage"[7] did not reveal any significant correlations with variation in the dependent variables. The other variables are discussed below.

3.3.2.1 Community

Although all variants are present in both Kalona and Holmes County, the frequency of occurrence differs between the two communities. The ranking of variants from most frequent to least is:

Holmes County, Ohio	[ɛ] > [æ] > [aɪ] > [e]
Kalona, Iowa	[aɪ] > [æ] > [ɛ] > [e]

While Holmes County speakers most frequently produce an advanced form, [ɛ], Kalona speakers favor the conservative form. The single speaker from Pennsylvania produced only [aɪ] tokens.

In terms of the borrowed English tokens, the Kalona speakers almost categorically retain the canonical variant [aɪ] (95%), while Holmes County speakers do so in only 52% of possible cases. This finding suggests that the restriction on incorporating English lexical items into PG phonology is much stronger in Kalona than in Holmes County.

3.3.2.2 Age

Age is strongly negatively correlated with the production of advanced variants. For age against vowel height, the Pearson correlation coefficient is $r = -.462$ and the $r^2 = .214$, which means that over 21% of the variation in vowel height can be accounted for by variation in age (and this despite the fact that vowel height as coded in this study is not truly a scalar numeric variable). Speakers over the age of 60 produce [aɪ] in two-thirds of their tokens. For speakers under the age of 60, the average frequency of [aɪ] tokens is less than 20%. Also speakers under the age of 40 produce a disproportionate number of the very advanced tokens, e.g., [eə].

This pattern holds true for both Kalona and Holmes County, although in every age cohort, the Kalona speakers have fewer advanced tokens than their Holmes County counterparts. These data suggest that Kalona lags behind Holmes County, by perhaps a generation, in the advancement and adoption of this sound change. A larger sample of free conversation is needed to confirm this and to rule out the possibility that the translation tasks, which differed somewhat in the two communities (see section 3.1), did not restrict the lexical and segmental environments for the Kalona tokens.

---

[7] This is true for dative usage only after another independent variable, "age," is factored out. There was a positive correlation between production of a high percentage of dative forms and production of a high percentage of [aɪ] variants. But the real correlation here is between both of these linguistic variables and the social variable age, see section 3.3.2.2.

3.3.2.3 Sex

In general, women produce fewer conservative tokens and more advanced tokens than men, but this phenomenon is limited to the two middle age cohorts 21-40 yrs and 41-60 yrs. The oldest and youngest age cohorts show few gender-correlated differences. Most remarkable is the relatively high percentage (17%) of very advanced tokens, e.g., [e], produced by women in the 21-40 year-old cohort. No other age group female or male produces more than 9%.

3.3.2.4 Denomination

The variable denomination singles out the New Order Amish who have significantly higher percentages of conservative [aɪ] tokens (over 50%) as well as a high number of advanced [e] tokens (10%). This bimodal distribution appears to be the result of a data sample dichotomy among the NOA in which two groups predominated: old, male church leaders, and young, female, office workers.

Comparing the Old Order Amish across communities reveals that the youngest age cohort (0-20 years old) have identical patterns of high [ɛ] usage and low [aɪ] usage in both Kalona and Holmes County. There are significant differences in the older generations. In Holmes County the middle-age cohorts share the pattern of the youngest generation, while the over 60 generation differs dramatically with high [aɪ] usage. By contrast, in Kalona, there is steadily increasing usage of the conservative [aɪ] variant in each generation as age increases.

3.3.2.5 Social networks: church, work, family

The office worker network consisting of about fifteen persons (eight are represented in this study) working in two offices with considerable English customer contact produced significantly more advanced tokens for vowel height: over 60% were either [ɛ] or [e]. Since four of the eight office workers in this network study are Amish women in the 21-40 age group, it is possible that age, sex, and denominational factors interact with the network variable.

The comments of one speaker gave reason to expect a possible geographical network correlation. He noted the advanced ingliding tokens [eə] and [eə] ("almost like they put an extra vowel in there") and when asked what person or group of persons use these advanced variants, he identified a particular group of young women in the section of the factory that he formerly worked in. These women, he speculated, mostly came from the same area in the county. However, the church/geographical network results did not show such a correlation.

3.3.3 Summary of Variation

Although variation in the production of the of the phoneme /aɪ/ has linguistic correlates, the strength of the effect of the social variable "age of speaker" overwhelms these correlations as well as other social correlates. Regardless of phonetic environment,

the younger the speaker, the less likely it is that the conservative [aɪ] variant will be produced. Over half (52%) of the tokens of the most advanced variant, [eə], were produced by women in the 20-40 age group. In the following section I will analyze the potential for advanced tokens such as [eə] to effect phonemic change.

4 Incipient phonemic merger: production and perception

To evaluate the possibility of phonemic merger, we must first consider how the variation in /aɪ/ may produce tokens which overlap the vowel space of other PG phonemes. Although in terms of quality both short /æ/ and short /ɛ/ would appear to show some overlap with /aɪ/, the length difference of /aɪ/ appears salient enough to avoid mergers with these two vowels. A more likely candidate is the long vowel /eː/.

4.1 Commutation test

To test a potential merger of the phonemes /aɪ/ and /eː/, I created a commutation test (Labov 1994, 356). The corpus for the commutation test was fashioned by selecting the minimal pair /gaɪl/ 'horses' and /geːl/ 'yellow' and randomizing twelve occurrences of each word in a single list. This produced a single list of twenty-four words which a native speaker then recorded for me. Since few PG speakers read PG, I used pictures to elicit the words. Finally, twenty words from the list of minimal pairs were played back to the person who recorded them (beginning on the third token and ending on the twenty second token to help ensure that the listener did not memorize the order of recording) and the person was asked to identify which word she or he had said (i.e., either 'horses' or 'yellow') for each token. If the person is unable to do so above the level of chance (50%), then we have convincing evidence of (near) merger phenomena. The evidence from the commutation test is particularly compelling, since speakers rate their own speech from a highly focused task in which the fact that minimal pairs are being elicited is obvious.

A second commutation test was created using the minimal pair /saɪ/ 'pigs' and /seː/ 'sea.' Both of these commutation tests were administered to five Holmes County PG speakers. I selected speakers under the age of 40, since my earlier quantitative data showed them to be most likely to produce advanced /aɪ/ variants. The results are given below in Table 6.

Table 6. Percent correct on commutation tests

speaker/listener (age, sex, denomination)	/gaɪl/ 'horses' vs. /geːl/ 'yellow'	/saɪ/ 'pigs' vs. /seː/ 'sea.'	TOTAL % correct
1. 30, male, New Order Amish	20/20 100%	20/20 100%	40/40 100%
2. 31, female, New Order Amish	20/20 100%	20/20 100%	40/40 100%
3. 32, female, Beachy Amish	20/20 100%	20/20 100%	40/40 100%
4. 32, female, New Order Amish	19/20 95%	15/20 75%	34/40 85%
5. 16, male, Old Order Amish	20/20 100%	20/20 100%	40/40 100%

Four of the speakers correctly identified all forty of their utterances. Of interest here is the one speaker who did not: speaker #4. This 32-year-old New Order Amish woman works in the office of a woodworking factory and earlier conversations with her had given me the impression that she is among the most advanced in her production of /aɪ/. The results of the commutation test show that clearly there is significant overlap in the phonetic space comprising the phonemes /aɪ/ and /eː/ for speaker #4.

In the /gaɪl/ vs. /geːl/ test she misidentified one word, but for the /saɪ/ vs. /seː/ test she incorrectly identified five words. Given that random guessing should yield a 50% correct score, her score of 75% is strong indication that for her, these phonemes are nearly merged. Her mistakes, however, were not completely random. In each of her errors she misidentified an /aɪ/ token as /eː/.

4.2 Cross-checking and extending the results of the commutation test

4.2.1 Commutation test cross-check

In order to verify that speaker #4 did not simply have perceptual difficulties, I had five other speakers listen to speaker #4's commutation test tokens.

Table 7. Cross-check: Speaker #4 commutation test with other listeners

listener (age, sex, denomination)	/gaɪl/ 'horses' vs. /geːl/ 'yellow'	/saɪ/ 'pigs' vs. /seː/ 'sea.'	TOTAL
A. 37, male, Beachy Amish	17/19 89%	13/19 68%	30/38 79%
B. 31, male, Beachy Amish	19/20 95%	18/20 90%	37/40 93%
C. 64, female, Beachy Amish	18/20 90%	18/20 90%	36/40 90%
D. = Speaker #2 (see Table 6)	19/20 95%	20/20 100%	95/100 95%
E. 65, female, New Order Amish	17/20 85%	17/19 89%	34/39 87%
TOTAL other listeners	90/99 91%	86/98 88%	176/197 89%
TOTAL including Speaker #4	109/119 92%	101/118 86%	210/237 89%

The results in Table 7 confirm a (near) merger of these vowels in the production of Speaker #4. Listeners are good—but not perfect—at distinguishing Speaker #4's /aɪ/ vs. /eː/. A roughly equal number of /aɪ/ and /eː/ tokens were misidentified and errors were scattered across 9 of 20 tokens for /gaɪl/ vs. /geːl/ and across 13 of 20 tokens for /saɪ/ vs. /seː/.

As a control, four listeners (B, C, D, and E) also listened to Speaker #2's commutation test. This check yielded only one error: 159/160 (99%) correct.[8]

At least one listener, Listener A, commented that it was very difficult to distinguish /aɪ/ from /eː/ in speaker #4's speech and he seemed surprised by the difficulty. He maintained that most speakers would not overlap the two phonemes in this manner.

4.3 Minimal pair test

I also had listeners listen to fifteen tokens of minimal pairs taken from sentences spoken by speaker #4 and speaker #2. The sentences had been elicited in an earlier translation task.

Table 8. Minimal pairs extracted from translation tasks of Spkr #2 and Spkr #4

/ai/ word class		/e:/ word class	
me: mail	'more miles'	me me:l	'more flour'
vais	'white'	'[ix] ve:s	'[I] know'
main	'mine'	'[ix] me:n	'[I] mean'
sai vasɐ	'his water'	se: vasɐ	'sea water'
[ix] bais	'[I] bite'	[ix bin] be:s	'[I'm] angry"
drei	'three'	dre:	'curve'

For each of these fifteen tokens listeners were asked to indicate which word from the minimal pair they heard, e.g., "Did you hear *more miles* or *more flour* or something else?"

[8] Listener A listened only to several tokens from the commutation test of speaker #5 and had "no problem" correctly identifying the tokens.

Table 9. Minimal pairs test results: number correct/total

Listener	Speaker #2		Speaker #4		TOTAL	
A. 37, male, Beachy Amish	4/6	67%	8/9	89%	12/15	80%
B. 31, male, Beachy Amish	6/6	100%	6/9	67%	12/15	80%
C. 64, female, Beachy Amish	6/6	100%	4/9	44%	10/15	67%
D. = Speaker #2 (see Table 6)	6/6	100%	6/9	67%	12/15	80%
E. 65, fem., New Order Amish	1 /2	50%	2/2	100%	3 /4	75%
TOTAL	23/26	88%	26/38	68%	49/64	77%

Again listeners have difficulty distinguishing tokens produced by Speaker #4. Listeners do better—but are not perfect—at distinguishing tokens produced by Speaker #2. There was a pattern to listeners' errors: 10 of 15 mistakes are /eː/ misidentified as /aɪ/. Three words were misidentified three times each: /veːs/ '[I] know', /beːs/ 'mean', and /meːn/ '[I] mean.'

4.4 Production of nearly merged sounds: acoustic measures.

In near-merger phenomena two different vowels are produced in a manner which causes them to be perceptually identical or nearly so. Yet acoustically significant differences may remain. Faber and Di Paolo 1995 suggest first testing for significant differences across several acoustic dimensions, then, if necessary, considering all of these dimensions simultaneously.

For the tokens in the commutation tests of speaerks #2 and #4, formant measures were taken at early, mid, and late points in the vowel (roughly at 20%, 50%, and 80% through the duration of the vowel). The acoustic dimensions tested were duration of the entire vowel, F1, F2, and change in F1 and in F2 from midpoint to late point in vowel. The average formant tracks for both speakers are given in Figure 6 and Figure 7.

Figure 6

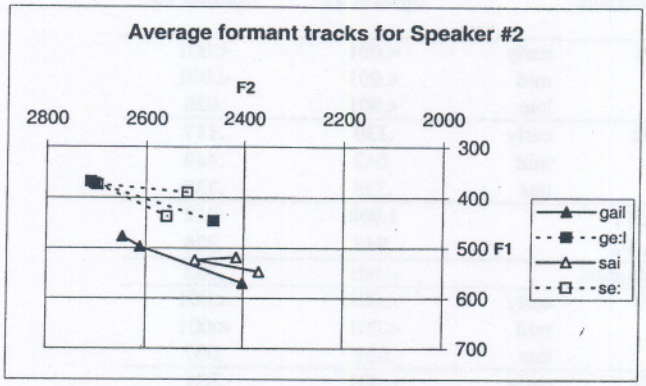
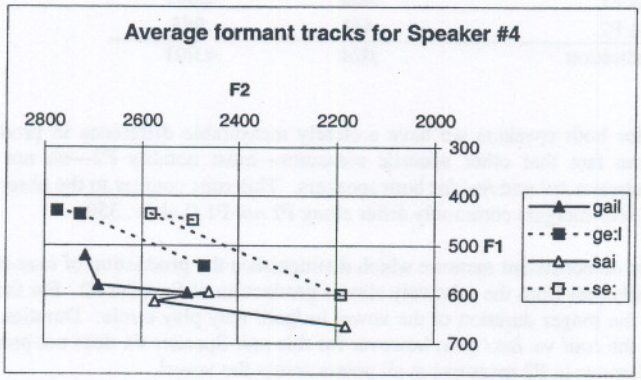


Figure 7



An ANOVA revealed a significantly higher F1 for /eɪ/ as opposed to /aɪ/ at all points in the vowel for both speakers. Measures of F1 of the same vowel across different words (i.e., /ge:l/ vs. /se:/) reveals no significant differences. (See Table 10 below in which only those differences which are <u>not</u> significant at <.05 are in **bold**).

Table 10. Tukey HSD post-hoc comparisons of ANOVA of commutation test

minimal pair	measure		Speaker #2	Speaker #4
/gail/ vs. /ge:l/	F1	early	<.001	<.001
		mid	<.001	<.001
		late	<.001	.026
	F2	early	**.130**	**.117**
		mid	.042	**.469**
		late	**.355**	**.735**
	Δ F1		**1.000**	.012
	Δ F2		**.845**	**.320**
	duration		<.001	**.332**
/sai/ vs /se:/	F1	early	<.001	<.001
		mid	<.001	<.001
		late	.001	.007
	F2	early	<.001	**.521**
		mid	<.001	**.985**
		late	.002	**.991**
	Δ F1		.008	<.001
	Δ F2		**.441**	**.995**
	duration		.024	<.001

Thus, for both speakers we have a clearly measurable difference in production (F1) despite the fact that other acoustic measures—most notably F2—do not differ significantly between /eː/ and /aɪ/ for both speakers. This runs counter to the observation that vowels in near-mergers commonly differ along F2 *not* F1 (Labov, 359).

There is no consistent measure which distinguishes the production of near-merged vowels by Speaker #4 from the relatively clearer production of Speaker #2. For the /gail/ vs. /ge:l/ pair, the longer duration of the vowel in /gail/ may play a role. Duration is not significant for the /sai/ vs. /se:/ pair, however for this pair Speaker #4 does not produce a significant difference in F2 measures at all points across the vowel.

The same basic pattern holds for the minimal pair test data: there are significant differences in F1 for /eː/ and /aɪ/ for both speakers at all points in the vowels. But duration and F2 do not differ significantly between /eː/ and /aɪ/ for both speakers. There is no clear acoustic cue to which we can attribute listeners' confusion on the minimal pair test.

4.5 Is phonemic merger underway?

A "near-merger" is defined as a contrast which speakers reliably produce but which they cannot reliably perceive (Labov 1994, 349-70). It is perception then, or rather the limits of perception, which drives the near merger process and the potential for

complete phonemic change. The results described in sections 4.1 and 4.2 clearly show that at least one speaker has partially merged the phonemes /aɪ/ and /eː/. Furthermore, native speaker-listeners were unable to consistently distinguish these phonemes in the speech of at least two Holmes County speakers. If we accept the relatively safe assumption that these two speakers are not unique in Holmes County, then we must also accept that continued spread of the advanced variants of /aɪ/ could lead to phonemic merger with /eː/.

Language contact may play a role in the retention or re-establishment of this phonemic contrast. English borrowings with /aɪ/ are resistant to monophthongization and raising, and are thus a constant source of renewal for the phoneme.

5 The origins of variation and change in the PG diphthong /aɪ/

Up to this point I have described the variable production of the PG diphthong /aɪ/, the linguistic and social conditioning that this variation is subject to, and the possibility of phonemic merger with /eː/. In this section, I will analyze four possible accounts for the introduction of this variation into the PG of Holmes County and Kalona. First I will consider two accounts based on dialect contact and language contact. Then I will consider two accounts based on motivations internal to PG.

5.1 External accounts: dialect contact or language contact

Dialects often differ in the phonetic details of a common phonemic inventory. When this is the case and speakers of the dialects are in contact with each other, it is possible that a particular dialectal variant will come to mark a particular sociolinguistic identity in a community and thus serve as a basis for change. Labov's study of variation in production of the American English diphthong /aɪ/ in Martha's Vineyard is a classic example.

There is some evidence to suggest that dialect borrowing/variation, at least at the lexical level, is already present in PG. The list of eight words in Table 11 are entered in Stine's 1990 dictionary as doublets having both /aɪ/ and /eː/ as possible pronunciations.

Table 11. Doublet entries in PG dictionary with /aɪ/ and /eː/ alternates

STINE listing	definition(s)	Modern German cognate
ʃaɪdə / ʃeːdə	to separate	scheiden
haɪling / heːling	cave (hollow)	Höhl-ung
laɪd / leːd	suffering (sorrow, mourning)	Leid
laɪʃt / leːʃt	molding, slat	Leiste
maiglix / meːxlix	probably	möglich
rais / reːs	journey	Reise
ʃvai / ʃveːgɐ	sister in law (brother in law)	Schwägerin (Schwager)
saine / seːnə	sift (strain)	seihen

All of the words in the above table reflect MHG vowels /eɪ/ and /œ/ which yield /eː/ regularly in PG. This then is not an example of the precise kind of dialect borrowing that we are looking for to describe the variation in /aɪ/[9], but it *is* evidence of dialect borrowing contributing to variation in PG.

In order to explore the possibility that dialect contact within PG might account for the current change in the vowel system, we must trace the development of PG /aɪ/ back to its MHG origins. Then we must examine the reflexes of these MHG vowels in the source dialects. It may be that several different reflexes of MHG vowels—reflecting the varied source dialect inputs to PG—have continued to co-exist in PG and thus have provided a model or target for the change of /aɪ/ to /æː/ or /ɛː/ or something else. The basis for such a model could simply be phonetic differences in production of the phoneme represented by /aɪ/.

5.1.1 Dialect contact: development of PG /aɪ/ word class from MHG and corresponding reflexes in PG source dialects.

The source dialects selected for comparison with PG in this study are: the Palatine dialect (Pfälzische) which is generally considered the most influential dialect in the genesis of PG[10], and two other dialects whose speakers are fairly well-represented among the early Anabaptist settlers in Pennsylvania and eventually Holmes County: Alsatian and Swiss, i.e. Low and High Alemannic.

Middle High German (approximately 13th century) provides the starting point for the development of these modern German dialects. Since the formative period for PG was approximately five hundred years later in colonial America (1683-1776), the relevant changes from MHG are those which took place between the 13th and the mid-18th centuries.[11] In most of the source dialects, the phonemic distinctions during this time period are fairly well understood, and have not changed considerably since that time period. Of course, the same cannot be said for the phonetic details[12], but we must make do with the imperfect and partial data that we have.

As noted in above, the PG /aɪ/ word-class comes primarily from diphthongization of the MHG long, high vowels /iː/ and /yː/. This change reflects similar changes in the Palatinate dialects. In the Alemannic dialects these MHG vowels remain monophthongs.

---

[9] The type of doublet that would be of most interest here is one involving MHG /iː/ having reflexes of both /aɪ/ and /eː/.

[10] See, e.g., Raith 1992, Reed 1972, Van Ness (1994, 421).

[11] German immigration to America resumed in the 19th century and a number of Amish and Mennonites came to America during that time. It is generally assumed that these later arrivals had little or no impact on the structure of PG. This may, in fact, be true for larger, older communities such as Holmes County. But in some of the smaller, newer communities, (e.g. Alsatians in Fulton County, NW Ohio and certainly the Swiss in Adams County, IN, see Thompson 1994) it may be that 19th century arrivals *did* leave some mark on the language, since they would have made up a sizeable minority or even majority in these settlements. The question of the impact of 19th century immigration will not be addressed in this paper.

[12] Russ (1982, 162) notes that the quality of the diphthong /aɪ/ can vary in current dialects from [ae] to [ei] and [ɛi].

The developments of MHG vowels in the source dialects and PG are summarized in Figure 8 below.

Figure 8. Development of MHG vowel /iː/ in non-Palatinate dialects and PG[13]

MHG	iː (includes merged yː)
Alsatian	iː, also ej in hiatus
Swiss: Berne	iː and yː (no merger); also ɛɪ in hiatus
PG and Palatinate	aɪ (and ɛː in Midwest PG); also oɪ in hiatus

Both Alsatian and Bernese Swiss retain the MHG monophthong, /iː/, and there is no direct model for Midwestern PG [ɛː] in either Alsatian or Bernese Swiss. Also, the range of variation within Midwestern PG includes monophthongal [æː] as well as diphthongal [aɪ] and [eə], but *no* speakers produce [iː]. Finally, although the Alemannic dialects both have diphthongal variants in hiatus position (defined by Keller as preceding a pause or a glide), this is precisely the position where PG also has undergone a different sound change the outcome of which does not figure into the discussion of /aɪ/. Lacking any further details of the phonetics of 18th century Alemannic and PG, it appears unlikely that Alsatian or Swiss dialectal influence has played a role in this change.

5.1.2 Contact with English

Holmes County PG speakers are in increasingly intense contact with English speakers some of whom speak a midland variety of American English in which the diphthong /aɪ/ is often monophthongized to low and slightly fronted [aː], e.g., 'right' pronounced as /raːt/. However, PG speakers overwhelmingly produce English words with diphthongal [aɪ] (see section 3.3.1.1), so contact with English, can be safely ruled out as a catalyst for monophthongization in PG.

5.2 Internal accounts: symmetry or drift

5.2.1 Restoring symmetry

Louden (1997) suggests that the monophthongization of /aɪ/ is internally motivated by an imbalance in the phonetic space of the long vowels in PG.

[13] The data on Alsatian and Bernese Swiss are taken from Keller 1961:125 and 92 respectively.

Figure 9. PG long vowels

i: u:
e: o:
ɔ:
a:

Louden (1997, 81) observes that the long vowel series includes three back, round vowels and only two front vowels (with /aː/ occupying a low central position). He claims this asymmetry is rectified by the monophthongization of /aɪ/ to /ɛː/ which then occupies a low front position opposite the back vowel /ɔː/ (the outcome of the monophthongization of the diphthong /aʊ/). This account rests on at least two assumptions: that oppositions based on length are salient in PG and that asymmetrical vowel spaces are inherently unstable.

As noted in section 2.1, there is ample evidence to suggest that, at least for the low vowels, distinctions based on length are crucial. However, the putative inherent instability of asymmetrical vowel spaces may be challenged on several counts.

Figure 10. Klamath (Penutian)

i
e o
a

Figure 11. Dialectal German

iː uː
eː oː
ɛː
aː

First, there are languages with unevenly distributed vowels, e.g., Klamath, a Penutian language which lacks high back round /u/ in opposition to /i/; also dialectal German, which has an asymmetry opposite that of Figure 9 in that it lacks a low back /ɔː/ as a counterpart to long front /ɛː/ (Hock:155). If even one generation of speakers acquires and maintains an asymmetrical system of this type, then we are obliged to reason that such a vowel system could exist as a stable system in any language for an indefinite period of time.

Second, languages with symmetrical vowel systems often undergo changes which eliminate the symmetry, e.g., Early Attic-Ionic which fronted the high back vowels resulting in a system with a three height contrast in the front vowels and only two in the back vowels (Hock:155).

Figure 12. pre-Attic-Ionic

iː	i	uː	u
eː	e	oː	o
ɛː		ɔː	
a	aː		

Figure 13. Early Attic-Ionic

iː	i	yː	y
eː	e	oː	o
ɛː		ɔː	
a	aː		

Still, a weakened version of Louden's argument still holds. That is, the arrangement of the articulatory and perceptual space for PG long vowels is such that there is a "vacancy" for an additional long, low, front vowel. While /aɪ/ is a likely candidate to fill this vacancy, it is certainly not obliged to do so by some principle of vowel space symmetry[14]. Probability is not the same as causation. Precisely what phoneme is most likely to fill this spot at a given point in time is subject to notions such as the naturalness of sound change and phonetic drift.

5.2.2 Drift

Low-level phonetic variation is a natural part of any language and can be heard in the speech of any one person at different points in time and between persons belonging to different social networks. Occasionally the cumulative nature of this variation across a speech community results in a phonetic change in a particular direction, a phenomenon Sapir labelled "drift" (1921:150, also Hock 1991:634).

In current continental German dialects, the phonetic realization of /aɪ/ can vary from [æ] to [ɛi] and [ɛi], and in Swabian (North Alemannic) variation can be seen in the orthography: *Zeit*, *Zait*, *Zoit*, *Zuit*, *Zäat*, 'time' (Noble 1983:62 and Russ 1982:162). The dynamic nature of the phonetic realizations of diphthongs in the dialects suggests that these diphthongs are subject to relatively rapid change internal to the system without any recourse to external pressures of dialect or language contact.

Furthermore, the direction of movement here—raising a long low vowel to a mid or high front vowel—has been observed in English, German, Greek, and Albanian, among other Indo-European languages (Labov 1994:116, 122). Another example of this type of change can be seen in the so-called secondary diphthongization in French where the putative change [ai] > [ɛ] occurs in such forms as Latin *lacte* 'milk' > [lait] (10th century) > [lɛt] (11th century).

In both the history of German and the history of English long vowels and diphthongs have undergone changes similar to the changes described for PG /aɪ/. For example, in southern American English /aɪ/ is produced as fronted [aː], which, in terms of

[14] I am not arguing here that maximum perceptual contrast between vowels, which often leads to a more or less symmetrical vowel space, is not a principle in the structuring of vowel systems (see Liljencrants and Lindblom 1972). I *am* arguing that there is no single optimal configuration which yields maximal perceptual contrast for a given vowel system. Indeed, Louden's account would be strengthened if it were framed in terms of perceptual contrast rather than "symmetry."

phonetic space, is not far removed from the PG variant [æː]. Both the Middle High German vowel shift and the Great Vowel Shift in Early Modern English involved the fronting and/or raising of long vowels (Labov 1994:124, 145).

Supporting evidence can also be found in studies of vowel coalescence. Cross-linguistic patterns of coalescence, the resolution of two adjacent vowels into a single vowel containing properties of both input vowels, demonstrate that sequences of low vowel + high front vowel (often across morpheme boundaries) are reflected in surface forms by the lowest front vowel in the language's inventory (Parkinson 1996:93-95). In PG this lowest front vowel could be either [æ] (a phoneme found primarily in English borrowings) or [ɛ]. These two vowels are the most frequently occurring monophthongs in Table 2, p.149.

What we observe, then, in Holmes County PG, appears to be change due to normal, internal variation in the language. Moreover, it is change of a relatively typical sort: the monophthongization and subsequent raising of the diphthong /aɪ/.

5.3 Spread of a sound change in PG

Within both the Holmes County and the Kalona communities, we see a sound change that is being led by the younger generations (see section 3.3.2.2). Within the younger generation in Holmes County, women who are employed in business offices appear to be leading the way in producing the most divergent variants. We can only speculate on the social motivations for doing so. Perhaps it is to mark oneself as "modern" within the constraints of Amish culture by speaking differently from "old-fashioned" PG speakers.

It is not clear whether young women have led the way throughout in the genesis and spread of this sound change. However, given that the economic opportunities afforded young women today are new to the community in the last part of this century, it seems unlikely that women in an earlier period would have had precisely the same social motivations. It is also unclear what social significance this variant in the speech of young women had/has in the wider community that would lead to it being adopted by others.

The quantitative data in section 3.3.2.1 suggest that the change of /aɪ/ from [aɪ] to [ɛː] is not proceeding at the same rate in Holmes County and Kalona. Furthermore, Louden suggests that the nature of the change differs substantially between the Midwest and Pennsylvania.

The changes in these three communities may share a common origin. If so, then we must account for how the change has spread from the community of origin to other communities. If not, then we must posit three parallel but independent changes. This latter hypothesis is certainly possible; however, given the striking similarities between especially the Midwestern communities, it seems more plausible to link the variation in Kalona and Holmes County as part of a single phenomenon. In the last section, I discuss

the ensuing difficulty in accounting for the spread of linguistic change between language islands.

6 Conclusion

I have presented data which confirms that a sound change which monophthongizes the PG diphthong /aɪ/—yielding /æː/ or /ɛː/—is in progress in two Midwestern Amish communities. Holmes County, Ohio speakers have advanced the sound change more than Kalona, Iowa speakers, but in both communities younger speakers (below age sixty) use monophthongal variants almost exclusively. The change is subject to some linguistic conditioning.

Perception experiments in Holmes County demonstrate that the most advanced tokens of this sound change, produced primarily by younger female office workers, are merging with the long mid-front vowel /eː/.

I suggest that this sound change is not the result of language or dialect contact or of system-balancing change, but rather, is simply an example of a relatively common type of language "drift."

This study delivers a proliferation of questions at its conclusion, among them:

- What is the nature of the spread, both in perception and production, of incipient phonemic merger or near-merger phenomena?
- Can the number of phonetic variants in the study be reduced from twelve (in Table 2) to just two or three that have clear sociolinguistic salience in the communities? How would this then change the patterns of variation?
- What can the restriction of this sound change to PG lexical items (vs. English) tell us about the (im)permeability of phonology in language contact and about the organization of phonology in code-switching and in the speech of bilinguals?
- What is the minimal level and means of interaction needed between dialect/language islands in order to maintain a high degree of linguistic homogeneity?

I will comment briefly on the last question.

The relative isolation of a speech community has long been recognized as a factor in both the development and maintenance of linguistic diversity. Conversely, geographic and social mobility have been understood as catalysts for the spread of changes and the eventual homogenization of dialects across a given region. Chambers (1995:66) calls the respective effects of isolation and mobility "natural linguistic laws."

While mobility and the resultant contact between speakers most often occurs between geographic neighbors, research in language and dialect contact has shown that interaction can occur between distant locales with little or no impact on intervening communities. Trudgill notes the spread of uvular /r/ between urban centers in Europe (1983:52,62) as well as the diffusion of the loss of /h/ from London to urban centers in East Anglia (1986:44-6). The homogeneity of African-American Vernacular English

across distant urban areas has also been noted in, among others, Fasold's study of the AAVE tense system comparing Washington D.C., New York, and Detroit (1972:219).

The primary difference between these examples and the study at hand is that, in the case of PG, we are studying *language* islands separated by regions inhabited by speakers of a different language, whereas the studies noted above (with the exception of the spread of uvular /r/ across dialects *and* languages) are concerned with *dialect* islands in which the intervening spaces are occupied by speakers of a mutually intelligible dialect of the same language. Still, the same principles should hold: mobility between islands will bring about homogeneity, isolation between the islands will encourage differentiation.

The fact that PG is "remarkably homogeneous" across geographical space (Van Ness:421) appears to be a violation of Chamber's "natural linguistic laws" of separation and mobility. Amish communtiies in the United States are scattered from Delaware to Montana, separated from each other by hundreds of miles and crucially lacking convenient access to modern means of transportation and communication. How have these apparently insular Amish communities—particularly in the Midwest—maintained a relatively uniform language, even down to the details of a particular sound change, for nearly a century and a half?[15]

Given, first, that the acquisition and spread of language generally occurs only via regular, face-to-face interactions between speakers and, second, that these Amish settlements have experienced nearly one-hundred fifty years of comparative geographic isolation, we would expect at least several dialects of Pennsylvania German to emerge (e.g., Ohio PG, Indiana PG, Iowa PG, etc. or rather Holmes County PG, Geauga County PG, etc.). The development of a relatively uniform Midwestern PG variety across these widely scattered speech islands remains something of an enigma.

If separation and mobility are indeed crucial factors or "laws" governing the spread of language change, then we are obliged to assume that these distant Amish communities are not as separated or immobile as they seem. They must interact in significant ways that are not visible to the newcomer. Multiple factors such as migration for economic, social (i.e., marriage), or religious (i.e., divisions and unions in church structure) purposes, visiting relatives, and even increased use of the telephone may play a role. Determining the precise nature of these interactions is a primary goal of future study.

References

Beam, C. Richard. 1991. *Revised Pennsylvania German Dictionary: English to Pennsylvania German*. Lancaster, PA: Brookshire Publications.

[15] What is more, the change in /aɪ/ is not the only parallel between the PG in Holmes County and Kalona. These communities pattern almost identically with respect to domains and degree of loss of dative case morphology across all age groups. See Keiser 1997.

Buffington, Albert F. 1939. Pennsylvania German; Its Relation to Other German Dialects. *American Speech* 14:276-286.

Buffington, Albert F. and Preston Barba. 1954. *A Pennsylvania German Grammar*. Allentown: Schlechters.

Chambers, J. K. 1995. *Sociolinguistic Theory*. Cambridge, MA: Blackwell.

Druckenbrod, Richard. 1994. *Mir Lanne Deitsch*.

Es Nei Teshtament. n.d. Committee for Translation. Sugarcreek, OH. South Holland, IL: The Bible League.

Faber, Alice and Marianna Di Paolo. 1995. The Discriminability of Nearly Merged Sounds. *Language Variation and Change* 7, 35-78.

Fasold, Ralph W. 1972. *Tense Marking in Black English*. Arlington, VA: Center for Applied Linguistics.

Frey, J. William. 1985 (1942). *A Simple Grammar of Pennsylvania Dutch*. Lancaster, PA: Brookshire.

Green, W.A. I. 1989. The Dialects of the Palatinate (Das Pfälzische). *The Dialects of Modern German*, ed. by Charles V.J. Russ, 241-264. Stanford: University Press.

Hock, Hans Henrich. 1991. *Principles of Historical Linguistics*. Berlin: Mouton de Gruyter.

Keller R. E. 1961. *German Dialects: Phonology and Morphology*. Manchester University Press.

Keiser, Steven Hartman. 1997. "From the Farm to the Factory: Divergence in Pennsylvania German?" presentation at New Ways of Analyzing Variation (NWAV 27). Athens, GA. October 2, 1998.

Labov, William. 1994. *Principles of Linguistic Change: Internal Factors*. Oxford/Cambridge: Blackwell.

Liljencrants, Johan and Björn Lindblom. 1972. Numerical Simulation of Vowel Quality Systems: the Role of Perceptual Contrast. *Language*. 48:4, 839-62.

Louden, Mark. 1997. Linguistic Structure and Sociolinguistic Identity in Pennsylvania German Society. *Languages and Lives: essays in honor of Werner Enninger*, ed. by James Dow and Michèle Wolff, 79-91. New York.: Peter Lang.

Meister Ferré, Barbara. 1994. *Stability and Change in the Pennsylvania German Dialect of an Old Order Amish Community in Lancaster County (Zeitschrift für Dialektologie und Linguistik:* Beihefte; H. 82). Stuttgart: Franz Steiner.

Noble, C.A.M. 1983. *Modern German Dialects*. Berne: Peter Lang.

Parkinson, Frederick. 1996. The Representation of Vowel Height in Phonology. Dissertation. The Ohio State University.

Penzl, Herbert. 1969. *Geschichtliche deutsche Lautlehre*. München: Max Hueber

Raith, Joachim. 1992. Dialect Mixing and/or Code Convergence: Pennsylvania German? *Diachronic Studies on the Languages of the Anabaptists*, ed. by Burridge & Enninger, 152-165. Bochum: Brockmeyer.

Reed, Carroll. 1972. A phonological history of PG. *Studies for Einar Haugen (Janua Linguarum* v. 59). 469-481. E. Firchow et. al. (eds). The Hague: Mouton.

Russ, Charles, V.J. 1982. *Studies in Historical German Phonology*. Bern: Peter Lang.

Russ, Charles, V. J. 1978. *Historical German Phonology and Morphology*. Oxford: Clarendon Press.

Sapir, Edward. 1921. *Language: An Introduction to the Study of Speech*. San Diego: Harcourt Brace Jovanovich.

Schlabach, Raymond. 1980. Some Phonological Aspects of the Pennsylvania German of Ohio. M.A. Thesis. The Ohio State University.

Stine, Eugene S. 1990. *Pennsylvania German to English Dictionary*.

Thompson, Chad. 1994. The Languages of the Amish of Allen County, Indiana: Multilingulaism and Convergence. *Anthropological Linguistics* 36:1.

Trudgill, Peter. 1983. *On Dialect*. Oxford: Basil Blackwell.

Trudgill, Peter. 1986. *Dialects in Contact*. Oxford: Basil Blackwell.

Van Ness, Silke. 1990. *Changes in an Obsolescing Language: Pennsylvania German in West Virginia*. Tübingen: Gunter Narr.

Van Ness, Silke. 1994. Pennsylvania German. *The Germanic Languages*, ed. by Ekkehard König and Johan van der Auwera, 420-38. London: Routledge.

Vella Laysa: Bivvel Shtoahris Fa Kinnah. 1997. Vella Dietsh, Millersburg, OH. Sugarcreek, OH: Schlabach Printers.

UNUS TESTIS, NULLUS TESTIS? THE SIGNIFICANCE OF A SINGLE TOKEN IN A PROBLEM OF LATER MEDIEVAL GREEK SYNTAX

Panayiotis A. Pappas

Abstract

In this brief paper I examine the placement of weak object pronouns in Later Medieval Greek when the verb is preceded by the negative marker οὐ. For the first time a detailed list of the occurrences of this phenomenon in 10 texts is presented and the distinction between οὐ "not" and ἄν οὐ "if not" is taken into consideration. The results show that pronouns are placed postverbally if οὐ precedes the verb, but preverbally if ἄν οὐ precedes the verb. I propose a tentative explanation for this differentiation based on the singular but robust occurrence of a counterexample in the same body of texts.

One of the more puzzling and under-examined phenomena of Later Medieval Greek syntax is the apparent variation concerning weak object pronoun[1] placement in the verb phrase[2]. The pronoun may appear either preverbally or postverbally as can be seen in examples (1) and (2) where both the verb and the element preceding the verb are the same, thus leaving us with no obvious explanation as to what causes the variation[3]. In the following pages I will present the results of an investigation into a well defined sub-area of

[1] Although Mackridge (1993, 1995) and Horrocks (1990, 1997) use the term *clitic* for these object pronouns I will refer to them as *weak (object) pronouns*, a theory-neutral term. I will refer to the string weak pronoun-verb or verb-weak pronoun as the *verb-pronoun complex*, while the elements that are believed to affect the ordering in this complex I will refer to as *environment*.

[2] Only the finite non-imperative verb forms are considered here.

[3] Mackridge's (1993, 1995) accounts are descriptive and do not capture any generalizations. Horrocks' (1990, 1997) accounts are given within the scope of much larger studies concerning the history of Greek and seem to have misinterpreted the data from this particular era (more on this below). Philippaki-Warburton's (1993) account lacks even descriptive adequacy (cf. Pappas forthcoming), while finally Rollo (1989) combines Byzantine and Cypriot Greek in his corpus, thus vitiating his analysis (see Mackridge 1993:326)

this problem[4], namely the position of the weak pronoun when the verb-pronoun complex is immediately preceded by the negative adverb οὐ (pronounced [u]).

(1)	πάλιν	λέγω	σας	
	palin	lεγo	sas	
	again	say-1sg pres	you-ACC WP	pl

"Again I say to you" (Digenis 1750)

(2)	πάλε	σας	λαλῶ
	pale	sas	lalo
	again	you-ACC pl WP	say-1sg pres

"Again I say to you" (Moreas, 715)

The particular problem of weak object pronoun placement in the environment of οὐ provides an ideal case for investigation as here alone do we find published disagreement about the facts concerning the variation. Horrocks (1990), while examining the placement of clitics (his word) throughout the history of Greek, wrote the following concerning οὐ and weak pronoun placement in Later Medieval Greek:

> ...the clitic was naturally drawn to second position within that complex[5], in accordance with the pattern we have seen many times already. This also tends to happen with the negative οὐ, which must similarly have been felt to "belong" to the verb in a particularly close way, both phonologically and semantically.

Although Horrocks is never explicit about it, I believe that the only way to interpret this statement is that Horrocks is identifying οὐ as one of the environments in which weak pronouns are placed in preverbal position. This is also evident from the example that he offers:

(3)	ἄν	οὐ	τὸν	εἰπῶ
	an	u	ton	ipo
	COND	NEG	he-ACC sg WP	say-1sg Perfective Pres

"If I do not say to him" (Ptochoprodromos III 43) (Horrocks (28))

On the other hand, Mackridge (1993:340), in his rule 1(b) makes the claim that "when the verb phrase[6] comes immediately after ... the negative adverb οὐ the order V+P is more or less obligatory" (cf example 4).

(4)	οὐκ	ἔμαθέν	το
	uk	εmaθε	to
	NEG	learn-3sg Perf Past	it-ACC sg WP

"He did not learn it" (Belissarios 269)

---

[4] The entire phenomenon is the subject of my upcoming dissertation *Weak object pronoun placement in Later Medieval Greek.*

[5] By *complex* Horrocks refers to the string complementizer (or negative marker)-verb, and not the weak pronoun-verb string as I do in this paper.

[6] Mackrdige's use of the term *verb phrase* should be equated to the term *verb pronoun complex* in this paper. It definitely does not refer to VP in standard syntactic theory.

What makes it especially difficult to assess the two contradictory statements is that the example offered by Horrocks is precisely the type of construction that Mackridge (1993:329) identifies as the only instance where the preverbal order is allowed:

> where οὐ coexists with ἄν[7] in the same clause, the pronoun is placed before the verb:
>
> (19) ἄν οὐ τὸ ἐπάρῃ
> an u to epari
> COND NEG it-ACC sg WP take-3sg Perfective Pres
> "If he does not take it" (Ptochoprodromos IV 514)

In Pappas (1997) I noted this discrepancy, but I felt I could not comment on it due to the small amount of data I had available at the time. Since thenI have expanded my database, and the results of a search concerning weak object pronoun placement in the environments of οὐ, ἄν οὐ, and ἄν can be seen in Table 1.

	οὐ		ἄν οὐ		ἄν	
TEXTS	P+V	V+P	P+V	V+P	P+V	V+P
DIGENIS	0	11	2	0	8	0
GLUKAS	0	4	0	0	0	0
PTOCHOPRODROMOS	0	6	3	0	10	0
SPANEAS	0	3	0	0	11	0
MOREAS (ln. 125-1630)	0	2	0	0	1	0
SPANOS (ms. D)	0	0	0	0	0	0
POULOLOGOS	0	0	3	1	2	0
BELISSARIOS (ms. N,V)	0	1	2	0	2	0
EROTOPAIGNIA	0	5	3	0	15	0
FALIEROS	0	0	0	0	1	0
TOTAL	0	32	13	1	50	0

Table 3. Variation of weak object pronoun placement in the environments of οὐ, ἄν οὐ, and ἄν

The detailed catalogue of the data shows clearly that Mackridge's evaluation was correct on both points. It does appear that when the verb-pronoun complex is immediately preceded by οὐ the pronoun is placed postverbally. On the other hand, when the conditional conjuction ἄν also precedes οὐ then the pronoun is placed preverbally in all instances but one. It seems then that Horrocks was indeed misled by the example he cited, or just the tokens with ἄν in general.

At the same time, even though Mackridge correctly identified the role that the preceding environments play in affecting the variation in weak pronoun placement his account is in essence descriptive and lacks explanatory force. The exact wording of his exception is (once again) that "where οὐ coexists with ἄν the pronoun is placed preverbally". Indeed, it is difficult to interpret the word "coexists" in any theoretical way. For instance does Mackridge use it as a synonym for "when ἄν precedes οὐ" i.e. for the case at hand?, or does he also mean "when οὐ precedes ἄν"?, a case that is not of any interest to us; after all this case is covered by Mackridge's (1993:340) rule 2: "the order V+P is more or less obligatory when the verb phrase is immediately preceded by the conditional conjuctions ἐάν, ἄν".

It seems to me that the most reasonable interpretation of Mackridge's statement is that he assumes that when ἄν and οὐ 'coexist', i.e. when they are placed side side before the

[7] pronounced [an].

verb pronoun complex, there is some type of formal conflict between the two affecting environments, which is resolved by postulating a rule precedence hierarchy in which the ἄν rule overrides the οὐ rule. Indeed in the (1995) paper, which is essentially the Greek version of the (1993) treatise Mackridge (1995:912) refers to conflict between rules 1(b) and 2: "Σε περιπτώσεις όπου συγκρούονται οι κανόνες (1β) και 2...". The verb συγκρούονται literally means to 'collide', thus validating the interpretation I offer of his (1993) term 'coexist'. Even this interpretation, however, runs into two problems.

First, both rules require that the affecting environment be *immediately* before the verb pronoun complex, so that in the ἄν οὐ cases there really is no *formal* conflict between the two rules, which means that the ἄν rule cannot override the οὐ rule.
The second problem, is that even if the rules were to be rewritten in order to accomodate cases like this one we would run into trouble in the case of the negative marker οὐ μὴ ([u mi]) where the pronoun is always placed preverbally (see Table 2, and examples (6, 7)).

TEXTS	P+V order	V+P order
DIGENIS	26	0
GLUKAS	11	0
PTOCHOPRODROMOS	13	1[8]
SPANEAS	28	0
MOREAS (ln. 125-1630)	4	0
SPANOS (ms. D)	7	0
POULOLOGOS	0	0
BELISSARIOS (ms. N,V)	5	0
EROTOPAIGNIA	21	0
FALIEROS	1	0
TOTAL	116	1

Table 4. Variation of weak object pronoun placement in the environment of μὴ

(6) οὐ μὴ σὲ βαρεθῶ
u mi sε varεθo
NEG NEG you-ACC sg WP be bored-1sg Perfective Pres
"(so that) I am not bored with you" (Poulologos, 366)

(7) οὐ μὴ τὰ γεύεσαι
u mi tα γευεsε
NEG NEG it-ACC pl WP taste-2sg Perfective Pres
"you do not taste them" (Ptochoprodromus, II 103)

Mackridge (1993:340—rule 2) also identifies μὴ as an environment associated with preverbal pronoun placement . Thus, what we have here is two cases where οὐ, an environment associated with postverbal pronoun placement is either preceded or followed

[8] This counterexample reads:

(5) Ἂν οὐν μὴ φθάση με
an un mi fθasi mε
COND so NEG suffice-3sg Perfective Pres you-ACC sg WP

"So if it is not enough for me" (Ptochoprodromos, I 271)

(immediately) by an environment associated with preverbal pronoun placement (ἄν and μὴ). In both cases the result is preverbal placement of the pronoun. Trying to sort this out with rule precedence arguments would fail since we would have to posit that in one case it is the environment closest to the verb that takes precedence (as in οὐ μὴ), while in another case it is the factor furthest from the verb that wins out (as in ἄν οὐ). Clearly this is not a desirable way to construct the grammar.

I believe, instead, that the solution may be found in the special status of οὐ. According to both Mackridge and Horrocks, οὐ seems to have been not an independent word, but rather a clitic. For instance Mackridge (1993:328) writes: "As for εἰ and οὐ they are clitics (i.e. unaccented), which may be the reason why they do not attract the object to the pre-verbal position" and we have already seen Horrocks' statement that οὐ "belonged" to the verb both phonologically and semantically. I interpret these statements to mean that οὐ and the verb form a single prosodic unit that either cannot be interrupted by the pronoun, or is simply considered as a verb form that has no independent word preceding it, so the pronoun has to appear postverbally according to Mackridge's (1993:340) rule 1(a) which states that "the order V+P is more or less obligatory when the verb phrase stands at the beginning of a clause" (cf. example 8):

(8)	'Αφῆκες	με	μνημόσυνον
	afikes	me	mnimosino
	leave-2sg Perf Past	I-ACC sg WP	memento-ACCsg

"You left me a memento" (Glukas 207)

I suggest that in the case of ἄν οὐ the negative marker is cliticizing onto the conditional conjuction ἄν instead of onto the verb. Although this position may be hard to substantiate without evidence from intonation and prosody (e.g. some clever way of looking at the metre of the lines in which ἄν οὐ appears), it is a more principled approach to explain the difference between the two cases than Mackridge's rule precedence argument for the reasons outlined above. Furthermore, by positing that the clitic negative marker οὐ could phonologically attach either to the following verb or the preceding ἄν we are also allowing for the logical possibility that even when ἄν is present the negative marker οὐ could still possibly attach to the verb creating the order ἄν οὐ Verb-Weak Pronoun, instead of the canonical ἄν οὐ Weak Pronoun-Verb. Such an example exists, as can be seen in table 1 and it reads:

(9)	ἄν	οὐκ	εἰπῶ	το
	an	uk	ipo	to
	COND	NEG	say-1sg Perfective Pres	it-ACC sg WP

"If I do not say it" (Poulologos 316)

Even though this type of example occurs only once in the database, and the traditional wisdom is that of *unus testus, nullus testus* I believe it would be wrong to dismiss this particular example on the basis that it is probably a scribal error, or any other kind of corruption of the original. First, there is the expert opinion of Tsavari (1987:89) who accepts the authenticity of this example and the fact that other editors (including Wagner, Zoras and Krawczynski) do not contest its authenticity either (cf. Tsavari (1987:222-239))prefering it over other textual traditions. Secondly, we need to take into consideration that we are not discussing a case of a unique counterexample to a well-attested construction. Instead we are dealing here with the rare possibility of an exception to a construction that is in itself so rare that in almost 12,000 lines of text there are only 12 occurrences of the norm. This number is so small that we cannot even build a statistically

valid sample for the construction (cf. Woods et al (1986)); from this numerical perspective the existence of a single token is indeed fortuitous. And third, this is not a case where we are left with a token for which no theoretical exegesis can be found. Instead we have a principled explanation that can account for this counterexample as a valid alternative construction based on the unique behavior of the negative adverb οὐ.

Thus, although my proposal that οὐ could attach either to a host before it (i.e. ἄν) or after it (i.e. the verb) is only a working hypothesis, I believe it offers the best promise for arriving at a full explanation of the insterestings facts of weak pronoun placement in the environment of οὐ and ἄν οὐ in Later Medieval Greek.

ACKNOWLEDGEMENTS

I would like to thank Brian Joseph and the participants of the Changelings reading group for their helpful comments and suggestions, and I hope that I have not misrepresented any of them in this paper.

REFERENCES

Horrocks, Geoffrey. 1990. Clitics in Greek. A diachronic review. In M. Roussou and S. Panteli (eds.) *Greek Outside Greece II*. Athens: Diaspora Books, pp. 35-52.

———. 1997. *Greek: A history of the language and its speakers*. London, New York: Longman Linguistics Library.

Mackridge, Peter. 1993. An editorial problem in the medieval greek texts: the position of the object clitic pronoun in the Escorial Digenes Akrites. In *Αρχές της νεοελληνικής λογοτεχνίας* Πρακτικά του δευτέρου διεθνούς συνεδρίου "Neograeca Medii Aevi" Βενετία 1993 pp. 325-342.

———. 1995. Η θέση του αδύνατου τύπου της προσωπικής αντωνυμίας στη Μεσαιωνική Δημώδη Ελληνική. *Μελέτες για την Ελληνική γλώσσα,* Πρακτικά της 15ης συνάντησης του Τομέα Γλωσσολογίας του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης 11-14 Μαίου 1994 (τιμητική προσφορά στον Καθηγητή Μ. Σετάτο.) Θεσσαλονίκη σσ. 906-929.

Pappas, Panayiotis A. 1997. Dispelling chaos: pronoun-verb placement in Later Medieval Greek: Paper presented at the 18th annual conference for Greek Linguistics; Aristotle University of Thessaloniki, Greece, April 1997.

———. forthcoming. *Weak object pronoun placement in Later Medieval Greek*. The Ohio State University, PhD. Dissertation.

Philippaki-Warburton Irene. 1995. Διαχρονική θεώρηση της θέσης των εγκλιτικών μέσα στην πρόταση. *Μελέτες για την Ελληνική γλώσσα,* Πρακτικά της 15ης συνάντησης του Τομέα Γλωσσολογίας του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης 11-14 Μαίου 1994 (τιμητική προσφορά στον Καθηγητή Μ. Σετάτο.) Θεσσαλονίκη σσ. 123-134.

Rollo, Antonio. 1989. L'uso dell'enclisi nel Greco Volgare dal XII al XVII secolo e la legge Tobler-Mussafia. In *Italoellinika* 2 pp. 135-146.

Tsavari, Isavella. 1987. *Ο Πουλολόγος*: Κριτική έκδοση με εισαγωγή, σχόλια και λεξιλόγιο Ισαβέλλας Τσαβάρη. Αθήνα Μορφωτικό ίδρυμα Εθνικής Τραπέζης 1987.

Woods, A., P. Fletcher, & A. Hughes. 1986. *Statistics in language studies*. Cambridge; New York: Cambridge University Press.

PRIMARY SOURCES

Βασίλειος Διγενής Ακρίτας (κατά το χειρόγραφο του Εσκοριάλ), και το άσμα του Αρμούρη. Κριτική έκδοση, εισαγωγή, σημειώσεις, γλωσσάριο, Στυλιανός Αλεξίου Αθήνα· Ερμής (1985).