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

1

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

2

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

3

%wL H*+L L%	H*+L L%	H*+L L%	H*+L L%
Yu'u-kun-to	Mine'yori-kun-wa	oni'isan-ni	aima'shita
2	2	2	3

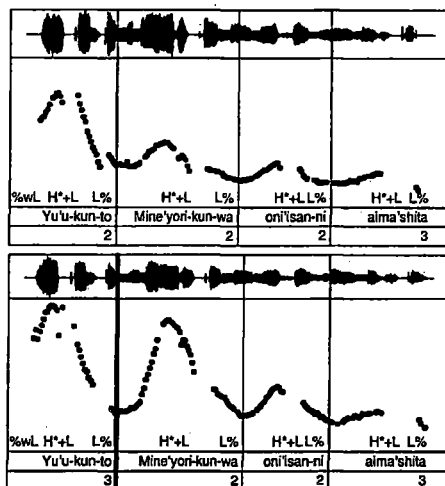| %wL H*+L L%	H*+L L%	H*+L L%	H*+L L%
Yu'u-kun-to	Mine'yori-kun-wa	oni'isan-ni	aima'shita
3	2	2	3

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

4

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.

5

paradigm was designed by Julia McGory and Stefanie Jannedy, and we are using it exten-
sively 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 ap-
propriately, 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 rela-
tive 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 par-
ticularly 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 informa-
tion. 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 utter-
ances 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
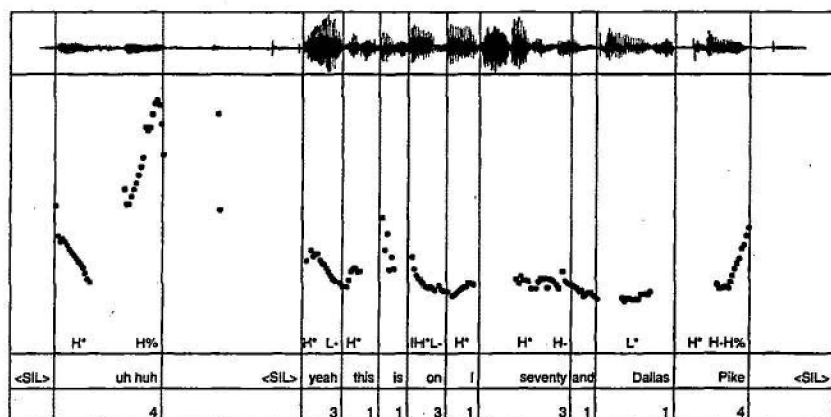(L+H*) pitch accent whose discourse function has been studied by Ladd [Ladd80], Ward

6

			H*	L-	H*		IH*L-	H*		H*	H-		L*		H*	H-H%	
H*		H%															
<SIL>	uh huh		<SIL>	yeah	this	is	on	I		seventy	and		Dallas		Pike		<SIL>
	4			3	1	1	3	1		3	1		1		4		

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
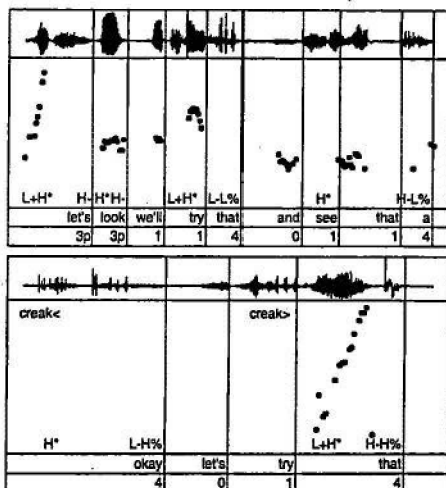
7

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
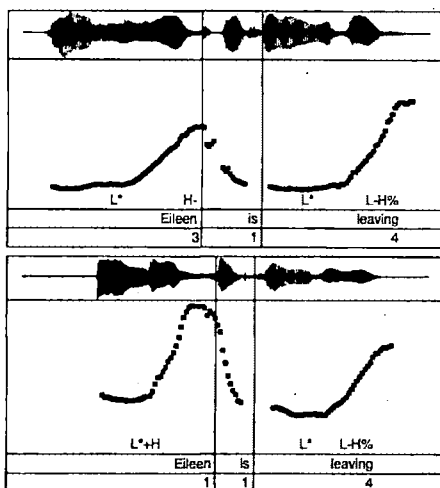
8

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 interpola-tion 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 ac-cent 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 Fig-ure 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

9

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.

10

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
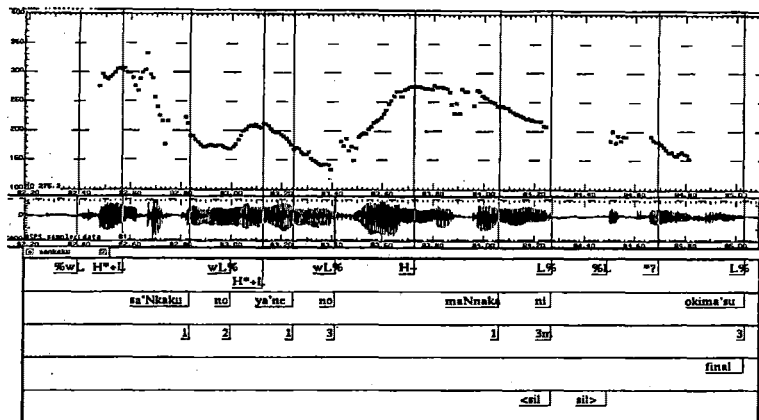
11

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 '!' 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
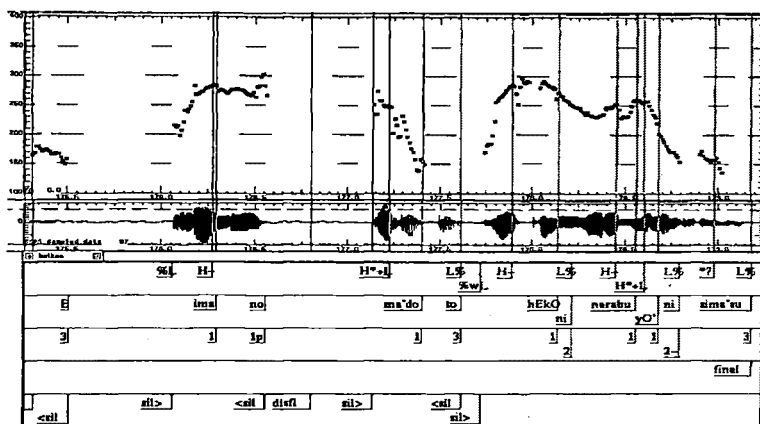
12

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

14

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].
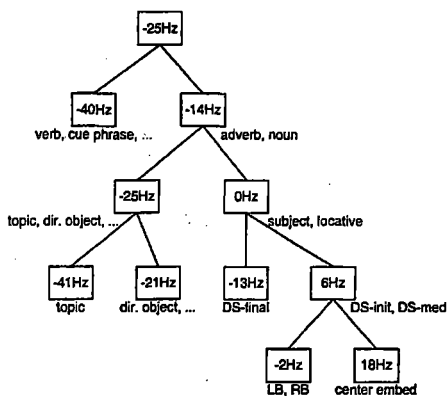
15

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.

16

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)

17

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
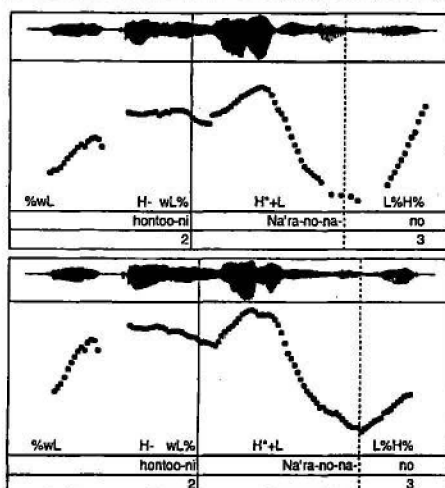
18

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

19

TAGGING PROSODY AND DISCOURSE

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

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.

22

[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 explicit-
             ness 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 Into-
             nation*. PhD thesis, Massachusetts Institute of Technology.

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

[PH90]   Pierrehumbert, Janet B. and Julia Hirschberg. 1990. The meaning of intona-
             tion 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 selec-
             tion 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 syn-
             thesis 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.