INTONATIONAL CUES TO DISCOURSE STRUCTURE IN JAPANESE

Jennifer J. Venditti† and Marc Swerts‡

†Linguistics, Ohio State University, 222 Oxley Hall, 1712 Neil Ave., Columbus, OH 43210 USA
‡Institute for Perception Research (IPO), PO Box 513, 5600 MB Eindhoven, The Netherlands

ABSTRACT

This study examines the extent to which intonation plays a role in the structuring of information in a Japanese monologue. The role of pitch accent in the intonational system of Japanese is very different from that in languages like Dutch or English: in Japanese, pitch accent is a lexical property of words and cannot be used to lend prominence to words at the sentence level. Therefore, we wondered if (and how) intonation can cue discourse structure in Japanese, comparable to how it is being used in Dutch and English. Results show that fundamental frequency (F_0), amplitude, and duration of the final accents in each sentence did not serve to cue the boundaries of discourse segments, contrary to our expectation. However, pitch range variations on NPs, examined in terms of their position in a discourse segment and their information status, did show a correlation with discourse structure.

1. INTRODUCTION

This paper examines the relationship between intonation and discourse structure in a Japanese monologue. A discourse is not a mere unordered string of sentences. Rather, sentences are grouped together into meaningful units, and each unit contributes to the overall communicative purpose of the discourse. Entities are introduced and referred back to as the discourse unfolds, constantly coming into and going out of the focus of attention. In other words, a coherent discourse typically exhibits a hierarchical structure in terms of major discourse segments, and is characterized by dynamic shifts in salience of discourse referents.

For many indo-european languages, it has been shown that intonational features such as F_0 range, F_0 movements, amplitude, speech rate, and timing can help cue the structure of spoken discourse (e.g. [4, 11, 1, inter alia]). More specifically, prosody may signal discourse structure in a dual way. First, in these languages, prosody can have a demarcative function, i.e., it can be used to highlight boundaries between major discourse segments. With respect to F_0 range, studies have shown that higher F_0 values mark the beginnings of discourse units, while the ends of such units are marked with a lowering of the F_0 range (e.g. [7, 6, 10]). Second, prosody can be used to mark which entities are salient to the discourse. Terken [12] found that in Dutch, expressions which refer to the topic of a discourse unit are most likely to have a pitch accent when first introduced. Nakatani [8] has found that accent distribution in English is closely related to the global and local salience of referring expressions.

It is relevant to check to what extent the results based on analyses of English and Dutch discourses can be generalized to Japanese. The latter language is different from the former in two important respects. First, Japanese has very explicit lexical marking of discourse structure: for instance, the morpheme *wa* signals that a specific NP is a topic. Put in simple terms, because of these explicit lexical markers, there may be less of a need to additionally mark topical structure by prosodic means. Second, the intonational system of Japanese is different from the systems of languages such as Dutch and English. In these languages, pitch accent is a property of the phrase, lending sentence-level prominence to words for some pragmatic effect. In Japanese, on the other hand, pitch accent is a lexical property of words, and it is possible for an intonation phrase to consist entirely of unaccented words. Thus, in Japanese, accentuation cannot be used to cue salient discourse entities in the same way that it does in Dutch and English. We hypothesize that variations in pitch range may be used for this purpose.

2. DATA

We based our prosodic and discourse analyses on a specific type of elicited spontaneous monologue. The experimental task was the same as that used in [12] for Dutch: speakers described how to construct the front view of a house using precut pieces of colored paper, providing enough information for a listener to reconstruct the house. Recordings of 4 native speakers of Tokyo Japanese (all female, age 25-40) were collected, and data from one speaker (KT) will be presented here. A fragment of her monologue, split into consecutive sentences, is given below:

1. tsugi-ni ee ie-no shoomengenkan-wo tsukemasu
   (*next erm house front door make*)

next, I will make the front door of the house

2. shoomengenkan-wa shiroi nagashikaku-no kami desu
   (*front door white rectangular paper*)
   I will use a white rectangular paper as the front door

3. kono shiroi nagashikaku-no kami-wo chairo-no shikaku-no ue-ni okimasu
   (*this white rectangular paper brown square on top place*)
   I place this white rectangular paper on top of the brown square

4. ichiban shita-no migigawa-ni shiroi doa-wo tsukemasu
   (*last down right side white door attach*)
   I attach a white door at the verry bottom on the right side

5. sorekara tsugi-ni yane-ni mado-wo tsukemasu
   (*then next roof window attach*)
   then I attach a window on the roof

From this example, it is clear that the experimental paradigm used here yields monologues which can be labelled in terms of discourse structure in a straightforward way. If the specific task structure is taken into account, discourse segments (DS) can be defined as the consecutive instructions, i.e. groups of sentences that deal with the same building block. Using this criterion, sentences 1 to 4 form one segment, sentence 5 starts the next one, etc. Topics can be operationalized as the noun phrases that refer to the different building blocks forming the central part of the consecutive instructions. In our example, the first segment deals with the front door (*shoomengenkan*), the next segment switches to a description of the window (*mado*), etc. The non-topical expressions are the other NPs within a DS, such as square (*shikaku*) in the first DS of the example. The monologue of speaker KT consists of 41 sentences. These can be grouped into 13 discourse segments, each conveying a certain purpose (i.e. to build one part of the house).

2.1. Discourse annotation

To check the reproducibility of our discourse analysis, 4 native speakers of Japanese were given the monologue text written in Japanese orthography (line breaks occurred at sentence boundaries) and were asked to divide the discourse into instruction units (defined as a unit of the discourse that describes a specific part of the house). Two of the judges had heard the speech file, while the other two had not. The discourse was divided into 13 units in exactly the same manner by all four judges. Topics and non-topics were assigned by the second author, and later checked by the first one.

2.2. Acoustic measurements

Using the Entropic Research Laboratory's Waves package, the monologues were prosodically labelled in terms of the Japanese ToBI transcription system [13]. Next, the data analysis was divided into two parts: first, in order to examine possible intonational cues to finality at the ends of discourse segments, we measured the average F_0, average rms amplitude, and duration of the syllable /ma/ in verb ending /-masu/. Since Japanese is a verb-final language, this accented morpheme occurs in absolute utterance-final position in each of the 34 of 41 sentences ending with a verb (others ended with either the copula or a different verb ending). This measure of average F_0 can be taken to indicate the local pitch range on the verb. In addition, in order to track the pitch range variation on referring expressions throughout the monologue, we took HiF0 measurements of each noun phrase in the discourse (in many cases in which the NP consisted of more than one intonation phrase, the highest F_0 value was used as the HiF0 measure of the NP).

3. RESULTS

3.1. General

Before embarking on the F_0 results, it has to be noted that there are obvious non-melodic cues to discourse structure. First, as is clear from the excerpt of the monologue given above, there appears to be clear lexical marking of discourse structure. There are, for instance, specific cue phrases to mark the beginnings of DS. Words such as *tsugi-ni* (next) and *sorekara* (then) were found at the beginning of 12 out of 13 segments. These phrases are generally low in the speaker's pitch range, and were always lower than the first NP of the segment. Of the 13 cue phrases which introduced segments concerning each piece of the house, 11 of these consisted of a single intonation phrase, with a H% boundary tone marked at its right edge; low pitch and separate prosodic phrasing have also been observed for cue phrases in English [5]. Additionally, there is explicit morphemic marking (e.g. *wa*) to signal discourse structure.

There are also other prosodic features than pitch to cue discourse structure. In addition to HiF0 relationships among noun phrases (see below), pauses also help to structure the discourse at the global level. A series of ANOVAs showed that pauses are significantly longer between discourse segments than within them, and within a DS pauses are longer between sentences than within them. This is similar to the result for Dutch in [11].

3.2. Specific

Cues to finality To check whether we could distinguish some phonetic differences among the utterance-final tones, we examined the effect of discourse position on F_0 of the accented syllable /ma/ in the verb at the end of each utterance, and also measured amplitude and duration. The hypothesis

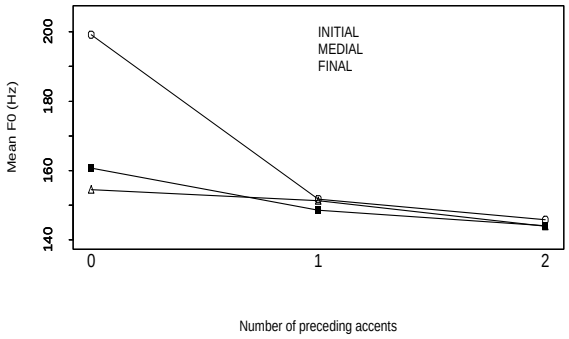

Figure 1: Interaction between number of preceding accents and discourse position on average F_0 values of /ma/.

was that there would be an effect of position on these acoustic measurements; namely, that the syllable would be lower in F_0 and amplitude and longer in duration at the end of discourse segments. Since in Japanese there is a phonological process of downstep whereby the local pitch range is reduced after an accent within the bounds of an intonation phrase, we also noted the number of lexical accents preceding /ma/ in the phrase, and included this as a factor in the analysis. A two-way ANOVA was conducted on each of the acoustic measures, using both accent and position as independent variables (position was divided into DS-initial, medial, or final, as determined by the four native judges). The first and last sentences of the monologue were excluded from the analysis because they were judged to be one-sentence DS. The results showed that there was no effect of position on any of the acoustic features measured, contrary to our hypothesis. However, there was a significant effect of accent on all of the measures (F_0: F=6.5, p=.006; AMP: F=3.4, p=.05; DUR: F=5.9, p=.008; df=(2,24) in all cases). The nature of this effect can be seen in the interaction between position and accent on F_0, shown in Figure 1.

This graph shows that the discourse position only matters when there are no preceding accents. In these unaccented cases, /ma/ syllables in DS-initial position are higher than those in either medial or final position. This interaction, though just a tendency (F(4,24)=2.2, p=.105), can be interpreted as follows: the preceding accent(s) reduces the pitch range to such an extent that the position distinction is lost. Only when there is no preceding accent can we observe the range of DS-initial verbs to be higher than in medial and final positions (which pattern together).

F_0 maxima The fact that there was no clear effect of discourse position on the scaling of final accents suggests that the pitch range falls to a common baseline at the end of each utterance. However, it is still possible that in other locations

	Topic	non-Topic
first mention	320 (48.11)	309 (44.76)
later mention	284 (38.72)	297 (38.26)

Table 1: Mean F_0 maxima (and standard deviations) of topic and non-topical referring expressions as a function of their first or later mentioning in the discourse segment

in the utterances a correlation between pitch range changes and discourse structure can be observed. To this end, the relationships among the HiF0 values of each noun phrase in the monologue were compared. The HiF0s of each NP (and of cue phrases) are plotted in Figure 2, which visualizes a portion of the monologue. F_0 maxima were analysed with respect to (i) position and (ii) information status.

Regarding position, a quick count reveals that the highest F_0 in each discourse segment usually occurs in the first sentence of a discourse segment (10/13 cases), which is comparable to the results of previous studies. In 5 of these 10 cases, the F_0 maxima in sentences appear to be higher than the maxima of the final sentences of preceding DS. Figure 2 gives an illustration of these phenomena: only the segment dealing with the livingroom window presents a counter-example.

Next, the F_0 maxima were examined as a function of the information status of the different NPs, differing between topical and non-topical referring expressions (see discussion of example above). In Dutch [12], as already mentioned in the introduction, it was found that the probability for an NP to be accented declines as a function of its serial position in a DS. That is, the probability of accentuation is high for first mention and thereafter remains low. In Dutch, this effect appears to be more marked for topical than for non-topical expressions. Terken [12] explains these results on the basis of coherence, i.e. high probabilities of accentuation after points of low coherence (after DS boundaries) and vice versa. The data in table 2 for our Japanese monologue are comparable in the sense that the serial position of an NP has an effect on pitch range. However, an unexpected finding is that topical expressions are higher than non-topical ones, only when they are first mentioned in a DS, suggesting that a more subtle discourse analysis is called for. Given this, we are currently investigating whether the F_0 relationships among NPs can be more elegantly explained within the framework of Grosz and Sidner [2, 3, 8, 9], which aims to model changes of the salience of entities at both a global and local level of discourse structure. First attempts indicate that this theory can indeed account for subtle difference in prominence relationships between NPs.

4. CONCLUSIONS

This study examined the role of intonation in cueing the structure of a spoken Japanese discourse. Specifically, we attempted to explain the pitch range variations in the discourse. We first found that there is a subtle effect of dis-

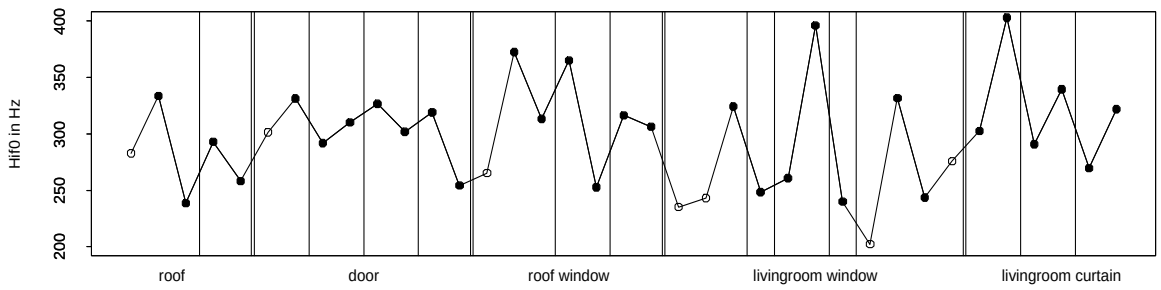

Figure 2: HiF0 values of consecutive NP (filled circles) and cue phrases (hollow circles). Double vertical lines show discourse segment boundaries and single lines show sentence boundaries.

course structure on the prosody of utterance-final syllables. The analysis of F_0 maxima reveals that both the position and the information status of an NP has an effect on pitch range. In the near future, we will extend the analysis to other speakers. We will also try to relate the range differences in Japanese on the basis of constraints previously used to explain the distribution of pitch accents in English in [8]. First observations suggest that, in two seemingly disparate intonation systems, these constraints work to relate prominence (be it cued by pitch accent or pitch range) to the dynamic shifts in the salience of entities across the discourse.

ACKNOWLEDGMENTS

We owe thanks to Mary Beckman, Jacques Terken, Kyoko Shimoda, and Shigemi Yamato for their helpful comments and insights. Part of this analysis was conducted while the first author was a visiting research assistant at ATR Interpreting Telecommunications Laboratories, Kyoto, Japan. Marc Swerts is also affiliated with the University of Antwerp (UIA).

REFERENCES

1. Ayers, G. 'Discourse functions of pitch range in spontaneous and read speech,' Ohio State University Working Papers in Linguistics 44: 1-49, 1994.
2. Grosz, B.J., Joshi, A.K., and Weinstein, S. 'Centering: A framework for modeling the local coherence of discourse,' Computational Linguistics 21: 203-225, 1995.
3. Grosz, B.J., and Sidner, C.L. 'Attention, intentions, and the structure of discourse,' Comp. Ling. 12: 175- 204, 1986
4. Hirschberg, J., and Grosz, B.J. 'Intonational features of local and global discourse structure,' Proc. of the Fifth DARPA Workshop on Speech and Natural Language, 441-446, 1992.
5. Hirschberg, J., and Litman, D. 'Now let's talk about now: Identifying cue phrases intonationally,' Proc. of the 24th Annual Meeting of the Assoc. for Comp. Ling., 163-171, 1987.
6. Hirschberg, J., and Pierrehumbert, J. 'The intonational structuring of discourse,' Proc. of the 24th Annual Meeting of the Assoc. for Computational Linguistics, 136-144, 1986.
7. Lehiste, I. 'The phonetic structure of paragraphs,' In Cohen, A., and Nooteboom, S.G., eds., Structure and Process in Speech Perception, Springer-Verlag, 195- 203, 1975.
8. Nakatani, C.H. 'Discourse structural constraints on accent in narrative,' To appear in van Santen, J.P.H., Sproat, R., Olive, J., and Hirschberg, J., eds., Progress in Speech Synthesis, Springer-Verlag, 1995.
9. Nakatani, C.H., Grosz, B.J., Ahn, D.D., and Hirschberg, J. 'Instructions for annotating discourses,' Tech. Report # TR-21-95, Center for Research in Computing Tech., Harvard Univ., 1995.
10. Silverman, K. 'The structure and processing of fundamental frequency contours,' Ph.D. thesis, Cambridge Univ., 1987.
11. Swerts, M., and Geluykens, R. 'Prosody as a marker of information flow in spoken discourse,' Language and Speech 37: 21-43, 1994.
12. Terken, J.M.B. 'The distribution of pitch accents in instructions as a function of discourse structure,' Language and Speech 27: 269-289, 1984.
13. Venditti, J.J. 'Japanese ToBI labelling guidelines,' Unpublished manuscript, Ohio State University, 1994.