PROSODIC MARKING OF INFORMATION STATUS IN TOKYO JAPANESE

Marc Swerts¹, Miki Taniguchi² and Yasuhiro Katagiri³

¹IPO, Center for User-System Interaction, Eindhoven, The Netherlands and CNTS, Center for Dutch Language and Speech, Antwerp, Belgium
²Graduate School of Language and Culture, Osaka Univ., Osaka, Japan
³Advanced Telecommunications Research Laboratories, MIC, Kyoto, Japan
m.g.j.swerts@tue.nl, t-miki@dab.hi-ho.ne.jp, katagiri@mic.atr.co.jp

ABSTRACT

This paper reports the results of a study on the prosodic marking of information status, i.e., whether information is ‘new’ or ‘given’ in the discourse, in Tokyo Japanese noun phrases (NPs). Previous investigations have indicated that such discoursal information is reflected in accent distribution in Dutch NPs, and in degree of accentuation in Italian NPs. An acoustic analysis of Japanese data, in which accent is lexically determined, reveals that information status, both for accented and unaccented words, is mainly reflected in pitch range, with higher range used to signal new information, and lower range used to signal given information. In addition to the effect of discourse, there is ‘downstep’, which lowers the pitch of words following an accented word inside a prosodic phrase. A perception study shows that this pitch range variation has limited cue value for listeners about the information status of particular words.

1. INTRODUCTION

This report describes research on prosodic correlates of information status in Tokyo Japanese noun phrases (NPs), which fits in a larger study on similarities and differences between languages in their use of prosody to signal contextual information. In Dutch and English, discourse context is reflected in the distribution of pitch accents in that such accents are used as pointers to information which is ‘new’ or ‘contrastive’ in the discourse, whereas ‘given’ information is typically deaccented. Support for such a correspondence comes from acoustic [7], perceptual [4] and cognitive [8] studies. These types of evidence may lead to the expectation that the relation between information status and (de-)accentuation is universal. Yet, there have been a number of studies showing that this is definitely not the case. In particular, Romance languages such as Catalan, Spanish or Italian, strongly resist deaccentuation, especially at the end of intonation groups. For instance, while Ladd [3] acknowledges that Italian allows deaccentuation on sentence level (e.g. repeated full NPs may be deaccented), he argues that Italian strongly disfavours deaccentuation within noun phrases (NPs) or other syntactic constituents.

Accent is a lexical feature in Japanese, so that in this language the accent status of words does not depend on the discourse context ([1], [3], [5]). For instance, the word “aoni” (blue) is always accented, whereas “kiiroi” (yellow) is unaccented. While this implies that, in Japanese, discoursal information cannot be signalled by accent distribution, it does not entail that such information cannot be cued by prosody at all. The literature ([3], [5]) mentions that differences in information status in Japanese are reflected in prosody in basically two ways, i.e., in phrasing and in prominence. On the one hand, there is the finding that information put in focus is produced as a separate prosodic phrase, whereas unfocused material is “dephrased” (as opposed to deaccented), meaning it is prosodically integrated in a larger phrase. On the other hand, it is reported, in line with observations for Italian (see above), that focused material is given a higher pitch than its unfocused counterpart, both for accented and unaccented words.

There are however a couple of factors which make the picture for Japanese more complex. First, there is the complicating issue of catathesis or downstep, defined as the phenomenon that “the H tone of a phonological word, whether accented or unaccented, will be lower if it follows an accented word than if it follows an unaccented word” ([3], p. 93; see also [5]). In principle, downstep is obligatorily triggered in such tonal sequences, but it is blocked by the occurrence of an intermediate phrase boundary. Second, there are words that are intrinsically more likely to become dephrased, and words which are more likely to become separate phrases. For instance, looking at a number of colour names, there appears to be a difference between words that represent “real” adjectives, such as “aoni” or “kiiroi”, and nouns that can become adjectives when the genitive marker -no is added, like “midorino” (green) or “mizuirono” (light blue). While this morpheme is obligatory for the latter type, it is optional for the “real” adjectives, so that “kiirono” and “aono” are also possible. Intuitively, words which lack a -no marker are more likely to be dephrased, whereas words with an (obligatory or optional) -no are more likely to form separate phrases.

Based on previous research, there are still a couple of unanswered questions regarding the prosodic marking of information status in Tokyo Japanese. First, most previous studies dealing with this issue are based on analyses of read-aloud data, the results of which do not necessarily generalize to spontaneous data. Second, there has been little research on the cue value of pitch range variation as a marker of information status. The goal of this paper is therefore to check whether some of the findings are also true for semi-spontaneous speech data, and whether pitch range variation is also perceptually relevant. In the current study, we will not focus on “dephrasing” and other prosodic features as possible markers of information status.
2. PRODUCTION STUDY

2.1. Method

The experimental materials were elicited by means of a variant of a paradigm used in previous research on Italian ([6]) and Dutch ([2]). Accent patterns for Tokyo Japanese were obtained in a (semi-) spontaneous way via a simple elicitation task. Subjects, eleven speakers in total (2 female, 9 male), were asked to describe stacks of differently coloured geometrical figures in an individual task, in such a way that it would be possible for a listener to reconstruct the sequence of figures from an available set on the basis of the recorded speech. The participants were somewhat constrained in the way they had to describe the figures: they were not allowed to use ellipsis (which may also cue information status; see e.g. [9]) and they had to describe the colour and figure names without uttering words like the Japanese equivalent of “and then” or “next” and without adding extra verbs like “put” or “take”. The data thus obtained allow an unambiguous operationalization of the relevant contexts. A property is defined to be new (N) to the discourse if it was not mentioned in the immediately preceding description, and it is given (G) if it was mentioned in the previous turn. This becomes clear by looking at the following example which gives a sequence of figure descriptions:

1. kiiroi maru  yellow circle
2. kiiroi shikaku  yellow square
3. midorino shikaku  green square
4. aoi sankaku  blue triangle

The geometrical figures and their respective colours were chosen such that the following conditions were met: (1) there had to be unaccented and accented versions of the figure terms and the colour descriptions and (2) the colour terms had to consist of nouns that need the genitive marker -no to become an adjective and ‘real’ adjectives where the use of -no is optional, while the presence or absence of -no did not change the accent status of the word. Also, the figure and colour names had to receive unanimous identification, which excluded the use of a colour name like “orange” since this can be translated into Japanese in two ways (“daidaiiro” or “orenzi”). Based on these criteria, the colour and figure descriptions given in figure 1 were selected for the experiment. The test materials also contained some dummy utterances, like descriptions of differently coloured circles (maru), that were not used for the actual analyses.

With these materials, different combinations of accented and unaccented colour (with -no optional or obligatory) and figure names were elicited in the description task, yielding for instance “midorino sankaku” (accented-accented) or “kiiroi shikaku” (unaccented-unaccented). By varying the sequential order of the figures, all possible combinations were collected for the eleven speakers in three different discourse contexts: (1) all new (NN), (2) only colour new (NG) and (3) only figure new (GN). The fourth logical possibility (GG; both colour and figure are given) was excluded1.

1In the following acoustic and perceptual analyses we will only concentrate on NG versus GN cases, given that the NN cases do not appear to behave in a uniform way, neither in terms of acoustics nor perception.
Speech utterances were recorded in an isolated studio, and then fed into the computer with a 16 kHz sampling frequency. The digitized audio data were analysed with the xwaves software package, including measurements of fundamental frequency ($F_0$). We measured the difference between $F_0$ maxima in the “colour” and “figure” names to see whether the discourse context influenced pitch range. In addition, we also looked at the average $F_0$ maxima in the “figure” names as a function of the accent status of the preceding word to test the effect of downstep.

2.2. Results

Figure 2 gives the difference between the $F_0$ maxima in the first and second word of elicited adjective-noun sequences for the two discourse contexts (NG and GN): negative values below 0 mean that the maximum in the second word is lower than that of the preceding word, and vice versa. Not surprisingly, it appears that these $F_0$ values in general reflect the accent status of the word sequences\(^2\), with large negative values for accented-unaccented combinations (kiiroi/midorino shikaku), with large positive values for unaccented-accented combinations (kiiroi/mizuirono sankaku), and comparatively less extreme values for accented-accented combinations (kiiroi/midorino sankaku) and unaccented-unaccented combinations (kiiroi/mizuirono shikaku). Concentrating on the discourse effect on pitch range, it appears that the $F_0$ differences in the left (NG) and right (GN) bars in the different figures, are always in the expected direction: that is, the maximum is relatively high with respect to the preceding maximum in the GN condition, and relatively low in the NG condition, meaning that pitch range reflects the information status in these cases. Note that this is true for all combinations of accented and unaccented words, though the effect is very small for “mizuirono shikaku”. Though not visualized in Figure 2, we also found an effect of downstep or catathesis on the $F_0$ maxima in “shikaku” or “sankaku”. In line with what one would expect on the basis of previous literature, it appears that these maxima are relatively low when the preceding word is accented, and relatively high when it is unaccented. This effect is true both for accented and unaccented nouns, with one exception: the maxima for the NG condition when the figure-name is “sankaku”, preceded by “kiiroi” or “aoi”.

3. PERCEPTION STUDY

3.1. Method

In order to explore whether the pitch range variation described above has some cue value for listeners, we performed a listening test in which subjects basically had to guess what the discourse context was for a particular utterance. The procedure for the test was as follows: we used all the utterances (i.e. the AA, AU, UA and UU noun phrases, both the long and short versions, elicited in the NG and GN conditions) from one speaker who - according to our own judgments - appeared to best signal given and new information in his speech. The reason to focus on the “ideal” speaker first, was to find out whether listeners were in principle able to hear cues about context from the way utterances are realized. In an individually performed test, listeners were presented with pairs of utterances (e.g. two realizations of “midorino sankaku”, one elicited in a GN context, the other in NG context). The interval between two members of a pair was 200 ms. The task for the listener was to guess which of the two represented the utterance in which either the figure or the colour was new information. In one test, listeners were instructed to focus on the noun (the shape), in another test on the adjective (the colour). Eleven subjects, who were not aware about the purpose of the research, participated in the experiment. Five subjects first did the task with the focus on the noun, the other six started with the focus on the adjective. The presentation of the different pairs was fully randomized. Listeners were allowed to listen to a pair as often as they liked, but once they had made a judgment on a given pair, they could not change their decision. The task was a forced choice test and no feedback was given to the subjects about the correctness of their answers. To give listeners the opportunity to get acquainted with the task and the type of speech materials, the actual test was preceded by a training session using three pairs from another speaker. They were not given a hint about possible cues to which they could pay attention, but after they had performed the task, they were asked as to what sort of features they had used as basis for their classifications. The whole test was fully automated (except for the questions after the test) and took approximately ten minutes on average per person.

3.2. Results

In their answer to the questions asked after the test, most listeners complained about the difficulty of the task. Still, most of them indicated that in doing the classification they mainly payed attention to prosodic features, including pitch range variation. There appeared to be quite some variation between subjects as to their capacity to correctly classify the utterances. The specific results for the utterance classifications are given in Table 1. Since there seemed to be no effect of the listening conditions (i.e., in which subjects had to focus on either the noun or the adjective), and of the order of the task and of the utterance pairs, we collapsed all the classification results per subject. The numbers in the table make it clear that the classifi-

\(^2\)Note the differences in scale values on the y-axis.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Maximum score</th>
<th>AA</th>
<th>AU</th>
<th>UA</th>
<th>UU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME</td>
<td>8</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>TK</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>TM</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>SM</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>YK</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>KS</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>ST</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>YI</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>HK</td>
<td>8</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>IM</td>
<td>8</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>YS</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

| All       | 88%           | 56.8 | 75.0 | 68.2 | 46.6 |

Table 1: Specific results of perception test: correct classifications for AA, AU, UA and UU phrases; both results per subject and for all subjects combined.
ations were not equally simple for all the utterances. Whereas AA and UU yield basically random scores, the performance for both AU and UA phrases appears to be better. Since we only used the speech data of one speaker as stimulus materials for the listening test, we are not sure yet whether the differences in the performance for the different noun phrases is due to the particular prosodic realizations of this speaker or to intrinsic properties of particular combinations of accented and unaccented words. A possible explanation for the better scores for UA and AU phrases could be that these word combinations allow better to judge the relative prominence of words, which should be a good cue for information status, whereas the prominence judgments could be more difficult for sequences of words which have equal status in terms of accent pattern. Obviously, the task for the listeners was difficult because in the current experimental set-up, listeners only had limited resources for their decision on whether information was given or new. In actual spoken communication, people can rely on many more cues such as morpho-syntactic features.

4. DISCUSSION AND CONCLUSION

Summarizing the main results of the acoustic and perceptual analyses that have been carried out so far, it can be seen that pitch range is influenced both by the discourse context and downstep. These two factors appear to have an orthogonal, independent effect on the $F_0$ maxima. In this way, the findings seem to confirm earlier observations made by e.g. [5] based on read-aloud data. The discourse effect on pitch range resembles that of Italian from an earlier study ([6]), in which it was found that Italian speakers could not easily deaccent given words inside NPs: in a similar task as the one presented above, the Italian speakers always accented both the colour and figure terms irrespective of the discourse context; yet, the difference between new and given information was encoded prosodically by means of gradient differences in accents, with smaller pitch accents on given information, and larger accents on new or contrastive information. For both accented and unaccented words, the current study finds similar degree differences reflecting information status in Tokyo Japanese, though these are not all perceptually relevant.

These findings remind one of the remarks of Ladd [3] on universal tendencies in intonation and, more specifically, on the principles-and-parameters view on prosody: "The basic idea is that there are a number of 'parameters' or dimensions of variation on which languages can vary in specified ways; in many cases these parameters are said to have a 'default setting' which can be overridden in a particular language" (p. 295). In this way, one is able to distinguish surface symptoms (that may be language-specific) from deep structural effects (that are universal). Based on the observations on languages that are quite different from an intonational point of view, one could argue that they are similar in that they treat focused utterance materials differently from unfocused ones, in that the former are realized more emphatically than the latter. Yet, the actual parameter settings between the languages differ considerably: Dutch distinguishes accented from unaccented words to mark the difference, Italian cues the difference by means of degree differences between accented words, and Japanese uses pitch range to distinguish both unaccented and accented words that are in focus or not. Obviously, a truly universal view on prosodic correlates of information structure should relate these findings to other linguistic phenomena that may distinguish languages, and that may be used to signal discourse context, such as word order, morphemic markers and ellipsis.

There are a number of possible future studies that could be done. First, a number of further acoustic measurements could be carried out to see whether information status is reflected in still other prosodic features (like duration, pausing and loudness), also to investigate the effect of “dephrasing”. Second, it may be worthwhile to try and elicit the types of NPs in a dialogue context, in which we expect speakers will make more use of pitch range to contrast or correct information than in our monologue setting. Third, in order to find out whether the perceptual results generalize to other speakers, we should obviously do the perceptual test also with the data from more than one speaker, though for that purpose we may want to use production data elicited in a dialogue setting.

Acknowledgments

The research reported upon in this paper was carried out while Marc Swerts was a visiting researcher at ATR-MIC during Fall 1999 and Summer 2000. Thanks are due to Bob Ladd, Haruo Kubozono, Mits Ota, Emiel Krahmer, Yoshinori Sakane and Noriko Suzuki for discussions and various kinds of assistance. Marc Swerts is sponsored as a postdoctoral researcher by the Flemish Fund for Scientific Research (FWO-Flanders).

5. REFERENCES