F0 CORRELATES OF TOPIC AND SUBJECT IN SPONTANEOUS JAPANESE SPEECH

John Fry
Linguistics Dept. & CSLI
Stanford University
Stanford CA 94305-2150 USA
fry@csli.stanford.edu

ABSTRACT

This paper examines F0 correlates of morphologically marked grammatical functions, in particular topic and subject, in spontaneous Japanese speech. Our data consist of F0 measurements of 7,106 nouns in the CallHome Japanese corpus of telephone conversations [4]. We find that topics exhibit higher peak F0 than subjects, contradicting information-structure accounts which predict that topics, which refer to ‘old’ information, should be less prominent. However, we suggest that the style and genre of speech is an important factor in this regard.

1. INTRODUCTION

In Japanese, the grammatical function of a noun phrase (NP) in a sentence is indicated by a postpositional particle like -wa (topic), -ga (subject), -ni (indirect object), or -o (direct object). This is illustrated in the list of sentences in (1), based on utterance (1a) from our corpus of telephone conversations (cf. Figure 1).

(1) (a) kareshi-ga tsurete kuru?
    boyfriend-SUBJ take come
    ‘Will her boyfriend bring her?’

(b) kareshi-o tsurete kuru?
    boyfriend-OBJ take come
    ‘Will she bring her boyfriend?’

(c) kareshi-wa tsurete kuru?
    boyfriend-TOP take come
    ‘As for her boyfriend, will he bring her?’ OR,
    ‘As for her boyfriend, will she bring him?’

(d) kareshi tsurete kuru?
    boyfriend take come
    [Ambiguous between (1a), (1b), and (1c)]

Japanese topics are complicated by the fact that there seem to be two subtypes of the topic particle wa: the thematic and contrastive. According to Kuno [3], a number of properties distinguish these two functions. First, thematic wa is used to mark ‘old’ information that is familiar to both the speaker and hearer, while the contrastive wa marks information (old or new) which the speaker wishes to contrast against something else. In addition, NPs marked by thematic wa are generally restricted to the beginning of a sentence (a defining feature of themes in many languages [8]). Finally, NPs marked by contrastive wa are characterized by more prominent, ‘contrastive’ intonation.

For example, changing the particle of kareshi ‘boyfriend’ to -o instead of -ga, as in (1b), would make the boyfriend the object, rather than the subject, of the verb. Similarly, changing the particle to -wa, as in (1c), makes the boyfriend the grammatical topic of the sentence, i.e. what the sentence is ‘about’ (although (1c) is ambiguous as to whether the boyfriend is the agent or patient). Finally, (1d) illustrates the grammatical ambiguity that arises when the particles are omitted, as frequently happens in casual speech.

Determining F0 and other prosodic correlates of Japanese grammatical functions is of both practical and theoretical interest. On the practical side, such data can be used (i) to help improve the naturalness of text-to-speech synthesis, and (ii) to help speech understanding systems overcome ambiguities that arise when particles are omitted, as in (1d). On the theoretical side, F0 data shed empirical light on information-structure accounts which predict that topics, which generally refer to ‘old’ information, should re-
receive less prominent intonation than subjects, which often introduce ‘new’ information [1, 8, 9].

2. PREVIOUS WORK

In a laboratory study, Finn [2] measured $F_0$ differences between tokens of noun-\textit{wa} and noun-\textit{ga} produced by 11 Japanese speakers who read aloud from a list of 12 constructed example sentences.\footnote{Finn’s original set of speakers was evenly balanced with six men and six women, but the data from one male speaker had to be discarded.} In Finn’s study, examples of thematic \textit{wa} and contrastive \textit{wa} were identified and treated independently. For each token of noun-\textit{wa} and noun-\textit{ga}, Finn measured the peak $F_0$ of the noun and the valley (minimum $F_0$) of the following particle (\textit{wa} or \textit{ga}). The drop in $F_0$ was then calculated as \textit{peak} − \textit{valley}. The averages for Finn’s $F_0$ measurements over all tokens are listed in Table 1. As the table shows, the noun-\textit{wa} tokens exhibited higher peaks and lower valleys than the noun-\textit{ga} tokens in Finn’s experiment, with the noun-\textit{wa} drops significantly greater than the noun-\textit{ga} drops ($p < 0.01$). Finn suggested that the lower valleys are partly explained by the fact that Japanese topics tend to be followed by longer pause durations, which correlate inversely with pitch. The data in Table 1 also support the generalization noted earlier that NPs marked by contrastive \textit{wa} receive more prominent intonation compared with thematic \textit{wa}.

Unfortunately, Finn’s study did not control for a number of other factors that are known to influence $F_0$, which complicates the task of interpreting her results. For example, Finn did not consider $F_0$ \textit{downdrift}, the compression of the pitch range on accentual phrases as they occur later in an intonation phrase. She also did not control for the lexical pitch accents of individual words, which might have played a role given the small number of sentences in the study.

A more recent study by Beckman & Venditti [1] examined pitch range variation in a collection of spontaneous and read Japanese monologues. Speakers were asked to narrate a story about two girls meeting in the park, following sequences of hand-drawn pictures as prompts. Transcripts of these spontaneous monologues were subsequently read aloud by other speakers. The collected speech data were then segmented, annotated with prosodic, syntactic, and discourse structure tags, and finally analyzed for pitch range variation.

Figure 2 shows part of a Classification and Regression tree (CART) used by Beckman & Venditti to model variation in peak $F_0$ in one of their read monologues. The value at each node is the average difference between observed peak $F_0$ and the value predicted by their ‘default’ pitch range model. The default model consists of ‘purely phonetic’ speaker-specific variables such as the amount of reduction at each downstep and typical initial values for the pitch range topline and baseline. The CART thus serves to model the influence of non-phonetic factors such as syntactic category and the position of an NP within its discourse structure (DS) (Figure 2).

The most striking feature of Figure 2 is the low pitch range of \textit{wa}-marked topics, at 41Hz below the predicted value, compared to \textit{ga}-marked subjects, which were produced right at the predicted height. Although the CART in Figure 2 is for a single monologue, Beckman & Venditti suggest that this example is not unrepresentative. They attribute the low pitch range of \textit{wa}-marked topics to their ‘old’ information status within a discourse segment [1, 9].

In sum, the studies by Finn and by Beckman & Venditti seem to point to opposite conclusions about the intonation of topics compared with subjects—the former finding higher peaks for topics, the latter lower peaks. We will propose an explanation for this discrepancy in Section 5 after reporting our own results.

3. METHOD

We examined 7,106 nouns and particles in the \textbf{CallHome Japanese} (CHJ) corpus [4], a collection of digitized speech data and text transcriptions of 120 spontaneous, unscripted telephone conversations. Each CHJ transcript covers a contiguous five- or ten-minute segment from a recorded conversation between native Japanese speakers. The speech data are sampled at 8KHz and encoded in NIST format on two channels (one each for caller and callee). The 120 conversations contain a total of about 340,000 word/morpheme tokens, 12,000 unique word/morpheme types, and 38,515 speaker turns. For this study, we restricted our attention to single nouns followed by \textit{wa}, \textit{ga}, \textit{ni}, or \textit{o}. Adjectives, relative clauses, and non-head nouns within complex NPs were

<table>
<thead>
<tr>
<th>Token</th>
<th>Subtype</th>
<th>$n$</th>
<th>Peak</th>
<th>Valley</th>
<th>Drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun-\textit{wa}</td>
<td>thematic</td>
<td>77</td>
<td>257.3</td>
<td>183.7</td>
<td>73.6</td>
</tr>
<tr>
<td>noun-\textit{wa}</td>
<td>contrastive</td>
<td>33</td>
<td>267.0</td>
<td>185.6</td>
<td>70.9</td>
</tr>
<tr>
<td>noun-\textit{ga}</td>
<td>—</td>
<td>44</td>
<td>267.0</td>
<td>179.2</td>
<td>87.8</td>
</tr>
</tbody>
</table>

\textbf{Table 1: Average $F_0$ values (Hz) from Finn’s study [2]}
<table>
<thead>
<tr>
<th>Token</th>
<th>Sex</th>
<th>n</th>
<th>Peak</th>
<th>Mean</th>
<th>Valley</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun-wa</td>
<td>F</td>
<td>1382</td>
<td>286.0</td>
<td>244.0</td>
<td>199.9</td>
</tr>
<tr>
<td>(topic)</td>
<td>M</td>
<td>303</td>
<td>189.1</td>
<td>158.7</td>
<td>126.3</td>
</tr>
<tr>
<td>noun-ga</td>
<td>F</td>
<td>2010</td>
<td>270.8</td>
<td>231.6</td>
<td>197.5</td>
</tr>
<tr>
<td>(subject)</td>
<td>M</td>
<td>340</td>
<td>175.6</td>
<td>150.0</td>
<td>125.3</td>
</tr>
<tr>
<td>noun-ni</td>
<td>F</td>
<td>2082</td>
<td>269.0</td>
<td>231.3</td>
<td>200.4</td>
</tr>
<tr>
<td>(ind obj)</td>
<td>M</td>
<td>363</td>
<td>176.8</td>
<td>150.5</td>
<td>130.3</td>
</tr>
<tr>
<td>noun-o</td>
<td>F</td>
<td>543</td>
<td>264.1</td>
<td>228.6</td>
<td>200.9</td>
</tr>
<tr>
<td>(dir obj)</td>
<td>M</td>
<td>83</td>
<td>170.6</td>
<td>142.2</td>
<td>120.1</td>
</tr>
<tr>
<td>U-initial</td>
<td>F</td>
<td>743</td>
<td>298.5</td>
<td>251.4</td>
<td>207.2</td>
</tr>
<tr>
<td>noun-wa</td>
<td>M</td>
<td>175</td>
<td>194.9</td>
<td>161.9</td>
<td>126.8</td>
</tr>
<tr>
<td>U-initial</td>
<td>F</td>
<td>922</td>
<td>280.7</td>
<td>237.4</td>
<td>201.2</td>
</tr>
<tr>
<td>noun-ga</td>
<td>M</td>
<td>158</td>
<td>179.7</td>
<td>153.0</td>
<td>126.9</td>
</tr>
</tbody>
</table>

Table 2: Average $F_0$ measurements (Hz) over 7,106 nouns

N is the first noun to appear in the utterance, even if it is not the first word of the utterance. Utterance-initial status was considered for two reasons. The main reason was to control for $F_0$ downshift over the course of the utterance. Secondly, as noted in Section 1, the thematic wa, unlike the contrastive one, is generally restricted to the beginning of a sentence. We thus reasoned that an utterance-initial token of noun-wa would be more likely to represent a true thematic topic and to refer to ‘old’ information.

4. RESULTS

Table 2 lists the average peak, mean, and valley measurements over all 7,106 noun-particle tokens. The most notable result is the high peak $F_0$ values for topics as compared to subjects (as well as objects). This is graphed in Figure 3. The higher topic peaks are mainly interesting with respect to the utterance-initial topics and subjects, which as noted earlier are better controlled for $F_0$ downshift and for contrastiveness. As expected, the utterance-initial topics and subjects both exhibit considerably higher-than-average $F_0$ values across the board, reflecting the well-known phenomenon of $F_0$ declination over the course of an utterance. A $t$-test on the utterance-initial topics and subjects found the initial noun-wa peaks to be significantly higher than the initial noun-ga peaks for female speakers ($t(1663) = 4.69$, $p < 0.0001$). For the sparser male data, the initial topic peaks were also higher than the initial subject peaks, though less significantly ($t(331) = 2.26$, $p = 0.01$).

We also performed a repeated-measures ANOVA on the set of peak $F_0$ values of all 4,035 noun-ga and noun-wa tokens. The goal of this test was to control for speaker effects and to test the influence of grammatical function (topic or subject) and utterance-initial status, as well as the interaction of these two factors, on peak $F_0$. Because of the heavy preponderance of female speakers in our data, the sex of the speaker was also added to the model. The identity of the speaker was then nested within the sex factor as a random effect. The results of the ANOVA, listed in Table 3, show that both utterance-initial status and grammatical function were significant factors.
F 
peak

F

V enditti [1] that

These results would seem to contradict the finding of Beckman &

ies, and in particular to the
different methodologies of the three stud-

tied). However, it seems likely that the reason topics are not less prominent than subjects in spontaneous telephone speech may be that speakers are more likely to simply omit old information rather than realize it with less intonational prominence.

Two final points should be noted regarding the information status of Japanese NPs. First, NPs with missing particles (cf. example (1d) from Section 1) were excluded from our study, although such NPs are very common in casual conversation and might be strongly associated with new or old information. Second, it should be emphasized that the binary ‘old vs. new’ distinction is a rather rough-grained metric of information status. For this reason, more fine-grained distinctions have been developed [6] which invoke notions like inferable or bridging references—i.e. references to objects that are associated with, but not identical to, a previously mentioned entity (e.g. referring to the door after mentioning the house). So although Japanese speakers are more likely to simply omit truly ‘old’ (i.e. contextually given) information in casual, spontaneous speech, they may still be exploiting wa and go in order to convey more subtle distinctions of information givenness.

Acknowledgments I’m grateful to Yasuharu Den, Edward Flemming, Jennifer Venditti, and Maria Wolters for helpful discussions, and to Emi Suzuki for segmenting the speech data. Naturally, errors are mine alone. This work was supported in part by U.S. NSF grant BCS-0002646 and by a dissertation grant from the Institute for International Studies at Stanford University.

5. DISCUSSION

Our experiment, like Finn’s laboratory study [2], found that wa-
marked nouns exhibited higher peak $F_0$ than ga-marked nouns. These results would seem to contradict the finding of Beckman & Venditti [1] that wa-marked topics exhibited considerably lower peak $F_0$ than ga-marked subjects (at least in the monologue re-

ported).

However, it seems likely that the different results can be attributed to a large extent to the different methodologies of the three studies, and in particular to the types of speech that were studied. For example, Finn’s data consisted of constructed example sentences which were read in isolation, with no surrounding context. As a result, neither information status (new vs. old) nor discourse structure could affect the prosody of subjects and topics. On the other hand, Beckman & Venditti were able to carefully control for both of these factors, since their speakers were narrating stories whose characters and structure were given in advance. Beckman & Venditti point out that there is a great deal of pitch range vari-
ation within syntactic categories like wa-marked NPs, and their models are able to account for this variation in terms of discourse structure and information status [1].

The results reported in this paper are based on spontaneous telephone conversations, which tend to be more terse, less structured, less grammatical, and more ambiguous than planned monologues or task-oriented dialogues. In particular, in casual Japanese speech the ‘old’ information is generally omitted from an utter-
ance (‘zero-pronominalized’) whenever it is clear from context ([7], pp. 362–364). This phenomenon is illustrated by utterance (1a) from Section 1 (cf. Figure 1), in which the object of the verb ‘bring’ is not mentioned in the sentence since it is already clear from the conversational context. We speculate, therefore, that the reason topics are not less prominent than subjects in spontaneous telephone speech may be that speakers are more likely to simply omit old information rather than realize it with less intonational prominence.

6. REFERENCES


