This paper examines $F_0$ correlates of morphologically marked grammatical functions, in particular topic and subject, in spontaneous Japanese speech. Our data consist of $F_0$ measurements of 7,106 nouns in the CallHome Japanese corpus of telephone conversations [4]. We find that topics exhibit higher peak $F_0$ 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)]

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.

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.

Determining $F_0$ 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, $F_0$ data shed empirical light on information-structure accounts which predict that topics, which generally refer to ‘old’ information, should re-
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 wa-marked topics, at 41Hz below the predicted value, compared to go-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 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 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 wa, ga, ni, or o. Adjectives, relative clauses, and non-head nouns within complex NPs were

---

1. 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.
Table 2: Average $F_0$ measurements (Hz) over 7,106 nouns

<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 (top)</td>
<td>F</td>
<td>1382</td>
<td>286.0</td>
<td>244.0</td>
<td>199.9</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>303</td>
<td>189.1</td>
<td>158.7</td>
<td>126.3</td>
</tr>
<tr>
<td>noun-ga (subject)</td>
<td>F</td>
<td>2010</td>
<td>270.8</td>
<td>231.8</td>
<td>197.5</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>340</td>
<td>175.6</td>
<td>150.0</td>
<td>125.3</td>
</tr>
<tr>
<td>noun-ni (ind obj)</td>
<td>F</td>
<td>2082</td>
<td>269.0</td>
<td>231.3</td>
<td>200.4</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>363</td>
<td>176.8</td>
<td>150.5</td>
<td>130.3</td>
</tr>
<tr>
<td>noun-o (dir obj)</td>
<td>F</td>
<td>543</td>
<td>264.1</td>
<td>228.6</td>
<td>200.9</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>83</td>
<td>170.6</td>
<td>142.2</td>
<td>120.1</td>
</tr>
<tr>
<td>U-initial noun-wa</td>
<td>F</td>
<td>743</td>
<td>298.5</td>
<td>251.4</td>
<td>207.2</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>175</td>
<td>194.9</td>
<td>161.9</td>
<td>126.8</td>
</tr>
<tr>
<td>U-initial noun-ga</td>
<td>F</td>
<td>922</td>
<td>280.7</td>
<td>237.4</td>
<td>201.2</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>158</td>
<td>179.7</td>
<td>153.0</td>
<td>126.9</td>
</tr>
</tbody>
</table>

We used the ESPS/waves+ pitch tracking software to extract $F_0$ measurements at 10ms intervals over the entire corpus, using a 5-point median filter for smoothing. We then used the forced alignment function of HTK [10] to automatically align the speech data to phonetic transcriptions word by word. As each automatic alignment finished, we inspected the results manually and made corrections to the alignments of all nouns and following particles based on spectral cues displayed by ESPS/waves+. For example, Figure 1 illustrates the aligned pitch track, spectrogram, and phonetic transcription for a sentence from conversation 3a_2196.

We considered separately the caller and the callee of each of the 120 conversations, for a total of 240 speakers. To help control for speaker effects, we eliminated 18 of the 240 speakers from our study because they had handed the phone to another speaker (usually a child) at some point during the conversation, resulting in two different speakers recorded on the same track. The remaining 222 speakers comprised 181 women and 41 men.\(^2\)

Using the lexicon supplied with the corpus, we identified 7,277 instances of noun-wa, noun-ga, noun-ni, and noun-o in the conversations. We eliminated 171 tokens for which the speech software could detect no voicing in either the noun or the following particle. For each of the remaining 7,106 nouns we extracted the mean and peak $F_0$ values for that noun, excluding the following particle. In addition, following Finn [2], we extracted the valley, i.e. the minimum $F_0$ for the corresponding following particle (wa, ga, ni, or o).

For each token of noun-ga and noun-wa, we recorded whether that token was utterance-initial or not. Since many spoken utterances begin with exclamations or hesitations, we defined utterance-initial to refer to any token of N-ga or N-wa that

\(^2\)Unfortunately, the sex distribution in the CHJ corpus is highly unrepresentative. Of the 120 callers recorded, only 37 (31%) were male. Furthermore, 62% of the male callers and 93% of the female callers placed calls to women rather than to men.

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 $F_0$ downdrift 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.\(^3\) 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 $F_0$ downdrift 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 have a significant effect on peak $F_0$.

\(^3\)The relatively low frequency of noun-o tokens in Table 2 ($n = 626$) is consistent with the observation of Martin ([5], p. 50) that the object particle o is very frequently dropped in spoken Japanese.
The ANOVA also indicated no significant interaction between grammatical function and utterance-initial status. In other words, the $F_0$ “boost” that topics received over subjects did not depend on whether the topics were initial or not. One might have expected some interaction based on the following two generalizations noted in Section 1: (i) utterance-initial tokens of noun-\textit{wa} are more likely to be thematic than those which come later in the utterance, and (ii) nouns marked by the thematic \textit{wa} tend to receive less prominent pitch than those marked by contrastive \textit{wa}. In other words, one might have expected utterance-initial tokens of noun-\textit{wa} to receive a relatively smaller boost over subjects compared with the non-initial noun-\textit{wa} tokens, but this was not observed.

The ANOVA also indicated no significant interaction between grammatical function and utterance-initial status. In other words, the $F_0$ “boost” that topics received over subjects did not depend on whether the topics were initial or not. One might have expected some interaction based on the following two generalizations noted in Section 1: (i) utterance-initial tokens of noun-\textit{wa} are more likely to be thematic than those which come later in the utterance, and (ii) nouns marked by the thematic \textit{wa} tend to receive less prominent pitch than those marked by contrastive \textit{wa}. In other words, one might have expected utterance-initial tokens of noun-\textit{wa} to receive a relatively smaller boost over subjects compared with the non-initial noun-\textit{wa} tokens, but this was not observed.

5. DISCUSSION

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

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 variation \textit{within} syntactic categories like \textit{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 utterance (‘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.

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 \textit{the door} after mentioning \textit{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 \textit{wa} and \textit{ga} 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.

6. REFERENCES


