WORD REPETITIONS IN JAPANESE SPONTANEOUS SPEECH

Yasuharu Den
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayamacho, Ikoma, Nara, 630-0101 Japan
den@is.aist-nara.ac.jp

Herbert H. Clark
Department of Psychology
Stanford University
Jordan Hall, Stanford, CA 94305-2130 USA
herb@psych.stanford.edu

ABSTRACT

This paper examines several hypotheses based on a ‘strategic’ view of word repetitions in English. We test whether these hypotheses also apply to Japanese with its fundamentally different syntax. Analyses of 10 task-oriented Japanese dialogues reveal two effects. First, pauses are more frequent before and just after a word at a suspension of the speech than after a repetition of that word. Second, the first token of the repeated word is abnormally prolonged. These results support the ‘strategic’ view of repetitions. Speakers often suspend speaking after making a preliminary commitment to a constituent, but they prefer to produce that constituent with a continuous delivery. These findings suggest the generality of these strategies across languages.

1. INTRODUCTION

Spontaneous speech contains various disfluencies such as filled pauses, self-repairs, and repeated words. These disfluencies seem to reflect problems in speech production. When speakers cannot formulate an entire utterance at once, or when they change their minds about what to say, they may suspend their speech and produce filled pauses or replace words they have already produced.

One traditional view of speech disfluencies is that they are merely accidents which are beyond speakers’ control. For example, speakers may accidentally suspend a word for some unexpected reason and restart it from the beginning. Another view of disfluencies is that they are the results of certain strategies under speakers’ control (Levelt, 1989; Clark & Wasow, 1998). Speakers may produce a filled pause to signal to their addressees that they are having trouble in speech production, or to inform their addressees of the kind of trouble they have.

In this paper, following the work by Clark and Wasow (1998) on word repetitions in English, we push the latter doctrine further by investigating word repetitions in Japanese spontaneous speech in order to test the generality of the ‘strategic’ view across languages. More specifically, we examine several hypotheses proposed by Clark and Wasow (1998) with our Japanese data to see whether or not they apply to a language with a fundamentally different syntax from English.

We will focus on word repetitions like the following one:

(1) ano Ya= Yamaguchi-to Hiroshima arimasu-ka uh Ya= Yamaguchi-and Hiroshima be-POLITE-Q uh, are there Ya= Yamaguchi and Hiroshima?

In (1), the speaker suspended the word Yamaguchi (a place name) after the initial mora Ya, and then restarted it from the beginning, resulting in a word repetition Ya= Yamaguchi.1 In Japanese, first tokens in repetitions are frequently cut off in the middle, as in (1), but in a few cases, speakers produce an entire word before repeating it, like Yamaguchi Yamaguchi. Both cases will be considered in this study.

2. COMMIT-AND-RESTORE MODEL

Clark and Wasow (1998) proposed the commit-and-restore model of repeated words, that employs three hypotheses about the sources of repeats.

The complexity hypothesis: All other things being equal, the more complex a constituent, the more likely speakers are to suspend speaking after an initial commitment to it.

The continuity hypothesis: All other things being equal, speakers prefer to produce constituents with a continuous delivery.

The commitment hypothesis: Some initial commitments to constituents are preliminary, with speakers already expecting, at some level of processing, to suspend speaking immediately afterward.

Clark and Wasow (1998) provided pieces of evidence supporting these hypotheses based on their analysis of word repetitions in corpora of naturally occurring English conversations. First, speakers are more likely to repeat an initial the or a of an NP when the NP is complex than when it is simple, and they are more likely to repeat the or a when the NP is at the left edge of a larger constituent than when it is at the left edge of a smaller one, supporting the complexity hypothesis. Second, speakers are more likely to add a delay just before initiating a constituent than after its first word, and they are more likely to add a silent or filled pause just before the restart of the constituent than just after, supporting the continuity hypothesis. Third, speakers often produce the first token of a repetition as a ‘phonological orphan,’ which is phonologically separated from the following items and pronounced with an unreduced vowel and/or prolongation, as in

1The symbol ‘=’ is used to indicate a word cut-off.
In the present study, the continuity and the commitment hypotheses are tested with our Japanese data. Due to the difference of syntactic structures between Japanese and English, word repetitions in Japanese have different characteristics from word repetitions in English. First, most word repetitions in Japanese are repetitions of content words (Den, Ishizaki, & Haruki, 1997), as opposed to function words, which are frequently repeated in English. In Japanese, repetitions of function words alone, e.g., *Yamaguti-to -to*, are very rare, although immediate corrections of function words by another function words, e.g., *Yamaguti-ni -to*, are common (Den et al., 1997). Second, as a consequence of the first characteristic, Japanese has no typical lexical items for repetitions, like *the* or *a* in English. Third, in Japanese, first tokens in repetitions are frequently cut off in the middle, which is less frequent in English. In our Japanese data, repetitions involving word cut-off amount to over 60% of the data. This is mainly due to long durations of content words, which are frequently repeated in Japanese.

These characteristics of word repetitions are particular to Japanese, and they could bring us different results from Clark and Wasow’s (1998) study in English.

### 3. METHOD

#### Data

The corpus used in this study consists of 30 task-oriented dialogues by 60 different native speakers of Japanese (total running time = appx. 7.5 hours, number of words = 73,258). The task was collaborative route finding, in which two participants looking at slightly different railroad maps, each unseen by the other, were asked to find a connecting path from a given start to a given goal. The two maps were different in that some connections on one map were missing on the other and in that some station names were missing on either map. The participants had to find a path which was available on both maps. This setting induced the participants to exchange spontaneously and naturally utterances about connections and station names. All dialogues were digitized on a computer and transcribed with part-of-speech and disfluency annotations.

Table 1 shows the distribution of various types of speech disfluencies found in the corpus. The disfluency categories are FP: filled pauses (excluding editing expressions involved in repairs or repetitions), ART: articulation errors, SUB: substitution repairs, INS: insertion repairs, DEL: deletion repairs, REP: word repetitions, and HYB: combinations of more than one of ART, SUB, INS, or DEL (Shriberg, 1994).

<table>
<thead>
<tr>
<th></th>
<th>FP</th>
<th>ART</th>
<th>SUB</th>
<th>INS</th>
<th>DEL</th>
<th>REP</th>
<th>HYB</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>3050</td>
<td>29</td>
<td>336</td>
<td>138</td>
<td>264</td>
<td>580</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 1: Distribution of speech disfluencies in the corpus.

#### Content words, 2.4 per thousand, reported by Clark and Wasow (1998), although it is much smaller than the repeat rate for English function words, 25.2 per thousand. The high repeat rate indicates the high spontaneity of our speech data, justifying the use of our data as a basis for empirical investigation.

For the present study, the first 10 of the 30 dialogues were selected for further analysis. These included 11 male and 9 female speakers.

#### Measurement

To test the continuity and the commitment hypotheses, we counted the number of pauses around word repetitions and the durations of repeated tokens. The precise measurement is illustrated in Figure 1.

First, we counted pauses just before first tokens (P1), between first and second tokens (P2), and just after second tokens (P3). At P1 and P3, only silent pauses were considered, whereas at P2, filled pauses and editing expressions were also regarded as pauses with the corresponding durations. Silences, or fillers, of 60 msec or longer were counted as pauses. To precisely measure pause durations, we carefully aligned the repeated tokens and the surrounding words with the speech data based on spectral cues displayed by the speech analysis software ESPS/waves+.

Second, we measured the normalized durations of first tokens (R1s) and second tokens (R2s). The normalization was needed to compensate for the differences of speaking rate across speakers and of inherent phoneme length across items. For each speaker and for each item, we collected from the corpus the fluent versions of the same word by the same speaker, where a ‘fluent’ version means an occurrence of the word not involved in a disfluency of any type listed in Table 1. Then, we calculated the normalized duration as the duration of a repeated token divided by the mean of the durations of its fluent counterparts. When a repeated token was cut off in the middle, we used the duration of the corresponding region in its fluent counterpart for normalization. Note that by this normalization, the value of 1.0 indicates that the token was produced at the same speaking rate as its fluent counterpart. We also calculated the normalized durations for the final phonemes of R1s and R2s by identifying these phonemes based on spectral cues.

The normalization procedure for token durations and the threshold value, 60 msec, of pause durations were also used by other researchers (Plauché & Shriberg, 1999).

#### Figure 1: Measurement used in the analysis.

- Presence of a pause at P1, longer than 60 msec.
- Duration of Ri, normalized by the mean of the durations of its fluent counterparts by the same speaker.

#### Table 1: Distribution of speech disfluencies in the corpus.
Exceptions. The following cases were excluded from the analysis:

1. Repetitions involving verbs and/or auxiliary verbs, e.g., (2).
   (2) a  wakat-ta   wakat-ta
       oh understand-PAST understand-PAST
       oh, I see, I see.

   These are likely to be for emphasis rather than disfluencies.

2. Repetitions overlapping/overlapped with the other’s speech, e.g., (3).²
   (3) A: Toyama-no  *tonari-no eki
       Toyama-GEN next-GEN station
       the station next to Toyama
   B:  *hidari hidari
       left left
       left

   These were excluded for two reasons. First, repetitions at the beginning of overlapping turns, as in (3), seem to have some social interactional origin (Schegloff, 1987), which would be better considered separately from the perspective of disfluency. Second, it was difficult to perform precise word/phoneme alignment for overlapping/overlapped speech as the speech signals of the two participants in a dialogue were not recorded on separate channels in our corpus.

3. Repetitions having no fluent counterparts. These were excluded simply because the normalization procedure could not be applied.

Predictions. The predictions from the continuity and the commitment hypotheses are as follows:

The continuity hypothesis: Pauses at P3 are less frequent than those at P1 and at P2.

The commitment hypothesis: The normalized duration of R1 is longer than that of R2.

According to the continuity hypothesis, speakers prefer to produce constituents with a continuous delivery. They don’t like to add a delay before every word when they have trouble in formulating an entire phrase or utterance. Rather, they are likely to suspend speaking at some point in a constituent and restore continuity to their delivery of the constituent after they have formulated it well enough. Thus, we can expect pauses after restarts to be less frequent than pauses before or during suspensions.

According to the commitment hypotheses, on the other hand, speakers can make a preliminary commitment at the beginning of a major constituent, even when they are aware of its not having been well-formulated. They do so, being pressed by a temporal imperative; by initiating a constituent, even if prematurely, to inform their addressees that they are engaged in planning the constituent, they can escape from being heard, due to a long delay, as opting out, as confused, or as having nothing immediately to contribute. Preliminary commitments, being a signal to addressees, are marked as such by some linguistic means, e.g., non-standard pronunciation (Fox Tree & Clark, 1997) and/or prolongation (Shriberg, 1999). Although Japanese has no pronunciation variation such as *thit vs. *thuh in English, we can at least expect first tokens of repetitions to be prolonged, or to have longer normalized duration than second tokens.

4. RESULTS

Pauses at P1, P2, and P3. The frequencies of pauses at P1, P2, and P3 are shown in Figure 2. For those cases that have at least one pause at P1, P2, or P3 (N = 43), a Cochran’s Q test revealed a significant difference among the numbers of pauses at the three locations (Q = 30.76, p < .001). Multiple comparisons using the Ryan’s procedure showed that pauses were significantly less frequent at P3 than both at P1 and at P2 (ps < .001). No significant difference was found between the numbers of pauses at P1 and at P2. These results support the prediction from the continuity hypothesis.

Normalized Durations of R1 and R2. The mean normalized durations of R1s and R2s are shown in Figure 3. The mean normalized duration of R1s was significantly longer than that of R2s (t(59) = 5.21, p < .001). R2 had approximately the same duration as its fluent version (mean normalized duration = 1.04), and the difference between the two tokens was due particularly to the prolongation of R1, which was, on average, 1.6 times longer than its fluent version. These results support the prediction from the commitment hypothesis.

Moreover, the prolongation of R1 was found to be attributed mainly to the prolongation of the final phoneme. Figure 4 shows the mean normalized durations of the final phonemes of R1s and R2s. The mean normalized duration of the final phonemes of R1s was significantly longer than that of the final phonemes of R2s (t(58) = 6.55, p < .001).³ Again, the final phoneme of R2 had approximately the same duration as its fluent version (mean normalized duration = 1.06). The final phoneme of R1, on the other hand, was, on average, 1.86 times longer than its fluent version. The degree of extension was greater for the final phoneme (1.86) than for the whole token (1.6).

³One data was excluded from the analysis for the final phoneme, since there was a mismatch of the final phoneme between the R2 token and its fluent counterparts due to the devoicing which occurred at the end of the R2 token.

²The symbol ‘*’ indicates the location at which the speech of the two speakers begins to overlap with each other.
How do these preliminary commitments affect conversational interaction? Without preliminary commitments, speakers might be forced to put an intolerably long pause before a constituent when they have trouble in formulating it. This would give addressees uneasiness or discomfort. Preliminary commitments, on the other hand, announce speakers’ engagement in the production of the following material, making an otherwise intolerable pause permissible by addressees. In this way, speakers are allowed to spend as much time as they need to prepare the following material to be delivered continuously. What we have shown in this paper is that this strategic use of preliminary commitments is not specific to a particular language with particular linguistic characteristics, but is a general strategy, across languages, for speakers to deal with the problems of how to speak in a timely fashion and yet how to speak smoothly.

There are, however, still several points to be investigated. First, the evidence for the commitment hypothesis we have provided is not as strong as the one for the English case. Prolongation alone might originate from some other reason; for example, it might be observed whenever a word is cut off. In order to intensify our claim, we have to show that the prolongation of first tokens is particular to word repetitions, and other ‘strategic’ disfluencies, and is never observed in, for instance, error correction repairs, which cannot be seen as strategic. Second, there could be a difference between Japanese and English regarding which word demands the suspension of the speech. In English, the troublesome item is typically the word immediately after the repeated token, e.g., literature in thy the literature, whereas in Japanese, it can be the repeated token itself, e.g., Yamaguti in (1), or a word after it, e.g., Hiroshima in (1). The former case is likely to be an accident rather than a preliminary commitment. The latter case, on the other hand, requires the repeated token to be usable for a preliminary commitment and, hence, to be free, by itself, from processing difficulty both for the speaker and for the addressees. In this regard, we have to examine the relationship between word repetitions and the contexts in which they occur.

5. DISCUSSION

As predicted, pauses at P3 are less frequent than those at P1 and at P2, and the normalized duration of R1 is longer than that of R2. Moreover, the difference of the normalized durations between R1 and R2 was due to the prolongation of R1, particularly at the final phoneme. These results strongly suggest that the continuity and the commitment hypotheses apply to Japanese as well as English. Speakers prefer to suspend speaking after making a preliminary commitment to a constituent that has not been formulated well enough, and reproduce the constituent with a continuous delivery. These preliminary commitments are marked as such by means, at least, of prolongation.

Figure 3: Mean normalized durations of R1s and R2s. Error bars indicate the standard deviations.

Figure 4: Mean normalized durations of the final phonemes of R1s and R2s. Error bars indicate the standard deviations.

ACKNOWLEDGMENTS

This research was conducted during the first author’s stay at the Center for the Study of Language and Information (CSLI), Stanford University as a visiting scholar supported by the overseas research program of the Ministry of Education, Science, Sports and Culture of Japan. The first author would like to acknowledge all the benefits gained from CSLI and the department of Linguistics at Stanford University. He also thanks John Fry for proofreading and comments.

REFERENCES


