A Conversation-Analytic Annotation of Turn-Taking Behavior in Japanese Multi-Party Conversation and its Preliminary Analysis

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 644–652
Marseille, 11–16 May 2020

A Conversation-Analytic Annotation of Turn-Taking Behavior in Japanese Multi-Party Conversation and its Preliminary Analysis

Mika Enomoto[1], Yasuharu Den[2], Yuichi Ishimoto[3]
[1]School of Media Science, Tokyo University of Technology, Japan
[2]Graduate School of Humanities, Chiba University, Japan
[3]Center for Corpus Development, National Institute for Japanese Language and Linguistics, Japan
menomoto@stf.teu.ac.jp, den@chiba-u.jp, yishi@ninjal.ac.jp

Abstract

In this study, we propose a conversation-analytic annotation scheme for turn-taking behavior in multi-party conversations. The annotation scheme is motivated by a proposal of a proper model of turn-taking incorporating various ideas developed in the literature of conversation analysis. Our annotation consists of two sets of tags: the beginning and the ending type of the utterance. Focusing on the ending-type tags, in some cases combined with the beginning-type tags, we emphasize the importance of the distinction among four selection types: i) selecting other participant as next speaker, ii) not selecting next speaker but followed by a switch of the speakership, iii) not selecting next speaker and followed by a continuation of the speakership, and iv) being inside a multi-unit turn. Based on the annotation of Japanese multi-party conversations, we analyze how syntactic and prosodic features of utterances vary across the four selection types. The results show that the above four-way distinction is essential to account for the distributions of the syntactic and prosodic features, suggesting the insufficiency of previous turn-taking models that do not consider the distinction between i) and ii) or between ii) or iii).

Keywords: turn-taking, conversation-analytic annotation, Japanese multi-party conversation, prosodic and syntactic features

1. Introduction

Turn-taking is a regularly occurring phenomenon in everyday conversations. To utilize realistic human-robot and human-agent dialog systems usable in our daily life, it is necessary that robots and agents be able to realize natural turn-taking behavior in multi-party conversation. Nevertheless, modern systems in practical use, such as Apple Siri, Google Home, and Amazon Alexa, cannot smoothly take a turn like a human even in simple two-party conversations. While human participants can respond to the previous utterance without a noticeable gap (Sacks et al., 1974), the current systems usually wait for a pause after the end of the previous utterance to recognize that they can take a turn. Considering that more than two participants are often engaged in everyday conversations, a possible next speaker will not wait for a pause if (s)he wants to prevent the turn being taken by another participant. In order to realize a system with a smooth turn-taking capability, the system needs to take a turn immediately upon detecting the end of utterance.

Many studies have been conducted to model human turn-taking behavior. Some studies (Koiso et al., 1998; De Ruiter et al., 2006; Maier et al., 2017) pointed out the importance of syntactic cues, such as part of speech, around the end of the utterance. Other studies (Gravano and Hirschberg, 2009; Friedberg, 2011; Niebuhr et al., 2013; Zellers, 2013; Arsikere et al., 2015; Gravano et al., 2016; Brusco et al., 2017; Masumura et al., 2017) demonstrated the efficacy of prosodic features such as the fundamental frequency and the intensity in the acoustic signal. These studies, however, used pause-delimited units for modeling and only considered the distinction between cases involving a turn switch between two participants and cases where the same participant continues his/her turn after a pause, whether the model is on-line or not.

Koiso and Den (2010) pointed out the above problem and proposed a proper model with reference to the turn-taking system for conversation proposed in the literature on conversation analysis (CA) (Sacks et al., 1974). They formulated a turn-taking model consisting of two distinct tasks: i) discrimination between *completion* and *non-completion* of the utterance; and ii) discrimination between *switch* and *holding* of the speakership upon completion of the utterance.

In this study, we focus on the second task above. We propose an annotation scheme that can be used to realize such a proper model and conduct a preliminary analysis towards that model. In Section 2., we describe our model and related previous works. In Section 3., we propose our annotation scheme motivated by CA studies. In Section 4., we describe the data used in the current study and report on the results of the annotation. In Section 5., we conduct a preliminary analysis of our turn-taking model, showing the need for precise distinctions between several turn-taking patterns. In Section 6., we conclude the paper and discuss future plans.

2. A Proper Model of Turn-Taking

A solution to the problem stated in the previous section may be found in an influential work in the literature of CA by Sacks et al. (1974) on turn-taking system for conversation. They described a turn-taking system that consists of two sub-components: i) the turn-constructional component and ii) the turn-allocation component. The first component concerns with the construction of basic units of interaction to which turns are allotted, *turn-constructional units* (TCUs). The second component describes two ways of allocating a new turn to one party: (a) the current speaker's selecting a next speaker and (b) self-selection by the next speaker. One typical procedure for option (a) is to use the first part of an adjacent pair, e.g., a question or request, affiliated with an explicit technique to address that utterance to a particular

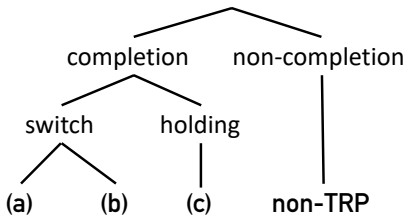

Figure 1: Our turn-taking model

co-participant, by, e.g., calling his/her name or directing the gaze toward that participant. Option (b) is for one participant to start a next utterance before anyone else. A set of rules is said to operate at every *possible completion* of a TCU, or *transition relevance place* (TRP). There are three options in the rule-set: (a) the turn is transferred to the next speaker by the use of a "current-speaker-selects-next" technique (procedure (a) above); (b) the turn is transferred to the next speaker by self-selection (procedure (b) above); and (c) the turn is continued by the current speaker. Following Koiso and Den (2010), the model can be straightforwardly depicted by the diagram in Figure 1, which represents a two-step discrimination of the turn-taking type.

Distinction between non-completion of the utterance at the first step and holding the speakership at the second step is very important because the risk of disrupting the flow of conversation when the system makes a wrong decision for turn-taking may be quite different in the two cases. This can be illustrated in the following examples.

```
(1) from Schegloff (1996, p. 74) (12)
    [simplified]
Marsha: Bu:t u-hu: his friend Steve and
        Brian er driving up. Right after::
        (0.2) school is out. And then he'll
        ^^^^^I
        drive do:wn here with the:m.
Tony:   Oh I see.
```

```
(2) from Sacks et al. (1974, p. 704) (a)
    in Footnote 14 [simplified]
Ava: He, he and Jo were like on the outs,
     you know?
     (0.7)
     ^^^^^II
Ava: [So uh,
Bee: [They always are ...
```

If the system erroneously decides to take a turn at point I in Example (1) and simultaneous talk by two participants occurs, this might disrupt the flow of conversation considerably because the system's utterance is recognized as harming the current turn. If, on the other hand, the system decides to take a turn at point II in Example (2) and simultaneous talk by two participants occurs, it might not be a great problem because such simultaneous talk is ubiquitous in human everyday conversations and there is a range of techniques for participants to deal with such a situation (Schegloff, 2000). Indeed, in Example (2), co-participant Bee starts his new turn simultaneously with the continuation of Ava's turn, but the overlap is soon resolved by declination of Ava's continuation.

Although Koiso and Den (2010) proposed this kind of model, they focused on the syntactic and prosodic features discriminating only between completion and non-completion of the utterance and did not investigate any features concerning turn allocation, i.e., discrimination between switch and holding of speakership.

More recently, Hara et al. (2019) independently pointed out the problem of the existing turn-taking models and proposed a two-stage computational model, similar to that in Figure 1, using deep learning, which can perform three-way discrimination among turn-switch, turn-holding, and non-completion in Figure 1. However, they did not distinguish options (a) and (b) in the turn-taking rules. The importance of distinguishing the two cases can be illustrated in the following example.

```
(3) from Sacks et al. (1974, p. 704) (a)
    in Footnote 13 [simplified]
Sara: Bill you want some?
                          ^III
Bill: No,
```

If the system (in role of Bill) erroneously decides to not take a turn at point III in Example (3) and a noticeable lapse occurs, it might be recognized as not attending to Sara's utterance or objecting to her offer. If, on the other hand, the system decides to not take a turn at point II in Example (2), this might not be a problem because the system has no obligation to take a next turn here and the previous speaker may continue his/her turn instead. Indeed, in Example (2), the previous speaker Ava tries to continue her turn.

Ishimoto et al. (2019) focused specifically on the distinction between options (a) and (b/c), i.e., whether or not a "current-speaker-selects-next" technique has been employed in the utterance. They also incorporated another type of units, i.e., *multi-unit turns* (Schegloff, 1996). A multi-unit turn consists of two or more TCUs that are projected to follow the prior part of the utterance to complete its content or the action accomplished therein. They are typically used in story telling or substantial explanation in which a single speaker exclusively holds a turn for a certain period of time. The following is an example.

```
(4) from Schegloff (1996, p. 61) (3)
    [simplified]
Ava: Oh my mother want to know how's your
     grandmother.
Bee: Uh:: (0.3) I don't know I guess she's
                              ^IV.a   ^IV.b
     aw- she's alright she went to the uh::
                        ^IV.c
     hospital again toda:y,
Ava: Mm-hm?
```

In Bee's turn, there are several points at which his TCU reaches a syntactically possible completion (points IV.a to c). These points, however, are not recognized as TRPs at which speaker-shift may be possible, and Ava eschews an attempt to take her own turn at those points. Although such recognition is based mainly on the initial part of the turn, i.e., "I don't know" in a position responding to a question, which is designed to project "more to come," Ishimoto et al. (2019) showed that there is a difference in prosody between possible completions of utterances within a multi-unit turn

and those of usual utterances. They failed, however, to find differences between options (a) and (b/c). If utterances regarding options (b) and (c) reveal different prosodic characteristics, the group (b/c) could be heterogeneous, which suggests that the distinction between (a) and (b/c) could be very vague. Therefore, the distinction between (b) and (c) is also necessary to properly understand the prosodic aspects of turn-taking behavior.

In this study, we focus on the syntactic and prosodic features that can distinguish the above four cases at a possible completion of an utterance: options (a), (b), and (c) and a multi-unit turn. In particular, we show that neither Hara et al. (2019)'s distinction between (a/b) and (c) nor Ishimoto et al. (2019)'s distinction between (a) and (b/c) is sufficient to precisely model our turn-taking behavior.

3. Annotation of Turn-Taking Behavior

3.1. Overview

In this section, we propose our annotation scheme for turn-taking behavior, which is motivated by CA studies described in the previous section. The annotation scheme is designed basically to distinguish options (a), (b), and (c) and multi-unit turns, but also to represent more fine-grained patterns observed in turn-taking behavior. More specifically, we employ two kinds of annotations, i.e., the *beginning* and *ending* types of the utterance, with reference to Schegloff (1996)'s idea of how TCU begins and ends.

3.2. Unit of Annotation

In the data used in this study (see 4.1.), speech has already been segmented into utterances based on *long utterance-units* (LUUs) (Den et al., 2010), which are regarded as a basic unit for interaction and determined considering syntactic, pragmatic, and interactional aspects. In most cases LUUs coincide with TCUs, but in some cases they are discrepant.[1] Some tags described below are designed to fill such a gap between LUUs and TCUs.

3.3. Ending-Type Tag

The ending-type tag represents how the utterance (LUU) ends. Figure 2 shows the taxonomy of the ending-type tags, and Table 1 provides a brief description of each tag.

First, irregular endings, occurring when the utterance is interrupted or abandoned in the middle, are labeled as `abort`. Second, regular endings are divided into two cases according to whether the utterance constitutes a substantial utterance or a response token (Den et al., 2011), the latter labeled as `mid-reaction` or `end-reaction`. Third, the endings of substantial utterances are classified into "possible completion" and "non-possible completion," the latter labeled as `mid-unit`. Note that `mid-unit` is used only in the limited circumstance that a turn-initial interjection is identified as a separate LUU but is better considered as composing a single TCU together with the following LUU; obviously, this does not cover all cases of non-TRP in Figure 1. Finally, "possible completions" are classified according to whether a "current-speaker-selects-next" technique has been employed or not, or the application of the turn-taking rules has been suspended within a multi-unit turn. These are labeled as `selecting`, `non-selecting`, and `multi-unit`, respectively. Note that the ending-type tag distinguishes only between options (a) and (b/c) in Figure 1 but not between (b) and (c).

[1]For instance, a turn-initial interjection is always treated as an independent LUU regardless of whether or not it constitutes an expectedly completed turn in the context, i.e., TCU.

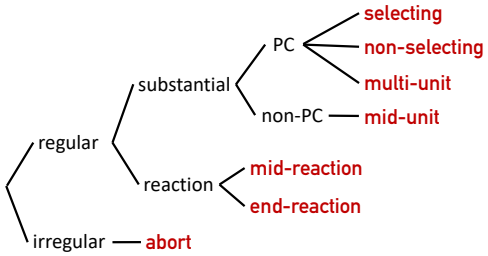

Figure 2: The taxonomy of the ending types of utterances

Table 1: The description of the ending type of utterances

Name	Description
selecting	Current speaker has selected next speaker
non-selecting	Current speaker has not selected next speaker
multi-unit	End of utterance within a multi-unit turn
mid-unit	After a turn-initial interjection
mid-reaction	To tie different sorts of response tokens
end-reaction	End of response token
abort	Interrupted or abandoned utterance

3.4. Beginning-Type Tag

In contrast to the ending-type tag, the beginning-type tag represents how the utterance (LUU) begins. Figure 3 shows the taxonomy of the beginning-type tags, and Table 2 provides a brief description of each tag.

First, irregular beginnings are labeled as either `early-start`, `late-start`, or `starting-over`. `early-start` is used when the utterance starts too early, i.e., earlier than the initiation of the predicate, which is placed in the end of an utterance in Japanese. `late-start` is used when the utterance starts too late (no strict time metric) and the flow of the conversation is disjoint; e.g., the second-pair part of an adjacency pair is delayed. `starting-over` is used when there is a lapse of more than 2 sec and the discourse topic has been terminated or broken before the utterance in question. Second, regular beginnings are divided into two cases according to whether the utterance constitutes a substantial utterance or a response token, the latter labeled as `begin-reaction` or `mid-reaction`. Third, the beginnings of substantial utterances are classified into "after possible completion" and "after non-possible completion," the latter labeled as `mid-unit`. Finally, "after possible completions" are classified according to various ways of starting a new utterance or extending the current utterance or turn; these include `other-selection`, `self-selection`, `continuation`, `increment`, and `multi-unit`.

Options (a), (b), and (c) in Figure 1 are represented by the following combinations of ending-type and beginning-type

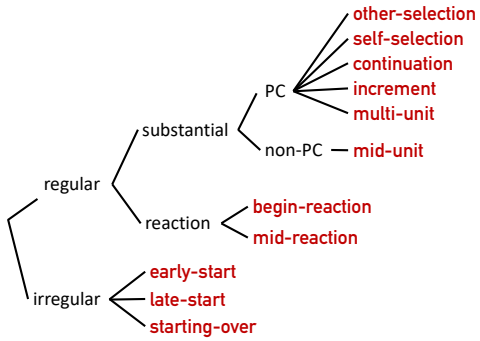

Figure 3: The taxonomy of the beginning types of utterances

Table 2: The description of the beginning types of utterances

Name	Description
other-selection	Selected as next speaker by previous speaker
self-selection	Self-selecting him/herself as next speaker
continuation	Current speaker continues the turn
increment	Current speaker adds a small grammatical unit to his/her possibly completed utterance
multi-unit	Beginning of utterance within a multi-unit turn
mid-unit	Following a turn-initial interjection
begin-reaction	Beginning of response token
mid-reaction	To tie different sorts of response tokens
early-start	Too early start of a new turn
late-start	Too late start of a new turn
starting-over	Restarting a topic or conversation

tags: (a) `selecting` followed by `other-selection`, (b) `non-selecting` followed by `self-selection`, and (c) `non-selecting` followed by `continuation`. There can be, however, different combinations, some of which may result in a deviant turn-taking behavior. In this study, we focus mainly on the ending-type tag.

4. Data and Annotation Summary

4.1. Corpus

For annotation and analysis, 12 conversations produced by 36 different speakers were selected from the *Chiba Three-Party Conversation Corpus* (Den and Enomoto, 2007) (Den and Enomoto, 2014). The Chiba corpus is a collection of casual conversations among three participants. The participants of each conversation were friends on campus. Each conversation was recorded using four digital video camera recorders and three headset microphones worn by individual participants. Each conversation was about 10 minutes long, and a total of 2 hours of conversations were used in this study.

The speech in the corpus has already been segmented into long utterance-units (LUUs), as described in 3.2., which we used as the unit of annotation and analysis.

4.2. Annotation Procedure

The first author, with good knowledge of CA, performed the annotation of the ending-type and the beginning type of all LUUs. The second author, who also has good knowledge of CA, then checked a part of the annotated data. When they did not agree on which tag to assign, the two annotators discussed the matter to reach a consensus. The disagreement typically occurred when the utterance, such as one ending with final particles *yo ne*, was ambiguous between assertion and clarification question; the ending-type tag of the first is `non-selecting`, while that of the second is `selecting`. In these cases, we decided to give precedence to `non-selecting`.

We have conducted no evaluation on annotation agreement, since our annotation relied heavily on our knowledge of CA rather than a written manual. In the future, however, we are planning to produce a publicly-available written manual for the research community, and will conduct agreement evaluation on other part of the corpus.

Table 3: The frequencies of the ending-type tags

Ending type	N	%
selecting	709	16.3%
non-selecting	2642	60.8%
multi-unit	399	9.2%
mid-unit	475	10.9%
mid-reaction	15	0.3%
end-reaction	5	0.1%
abort	101	2.3%

Table 4: The frequencies of the beginning-type tags

Beginning type	N	%
other-selection	554	8.8%
self-selection	2204	35.0%
continuation	1103	17.5%
increment	255	4.1%
multi-unit	369	5.9%
mid-unit	476	7.6%
begin-reaction	794	12.6%
mid-reaction	15	0.2%
early-start	248	3.9%
late-start	59	0.9%
starting-over	3	0.0%
begin-fragment	218	3.5%

4.3. Summary Statistics

The frequencies of the ending-type tags assigned to LUUs followed by one or more LUUs by the same or other participants are shown in Table 3. `non-selecting` was the most common tag (60.8%), followed by `selecting` (16.3%), `mid-unit` (10.9%), and `multi-unit` (9.2%). These substantial-utterance types amount to 97.2% of the entire data. The predominance of `non-selecting` would be a characteristic of casual conversations and is consistent with a general tendency that casual conversations contain a vast amount of statements compared with task-oriented dialogs (Shriberg et al., 1998).

The frequencies of the beginning-type tags assigned to LUUs following one or more LUUs by the same or other

participants are shown in Table 4. self-selection was the most common tag (35.0%), followed by continuation (17.5%), begin-reaction (12.6%), other-selection (8.8%), mid-unit (7.6%), multi-unit (5.9%), and increment (4.1%). Except for begin-reaction, these are among the substantial-utterance types, and amount to 78.8% of the entire data. The predominance of self-selection would be a natural consequence of the predominance of non-selecting in the ending-type tags. Note that begin-reaction occupies as much as 12.6% of the data, which reflects a frequent use of response tokens in Japanese conversations (Maynard, 1989).

5. Preliminary Analysis of Annotated Data

5.1. Purpose

In order to demonstrate the contribution of our annotation to the study of turn-taking models, we conducted a preliminary analysis of how the syntactic and prosodic features of the utterance vary depending on the ending type of the utterance. More specifically, we show that a binary distinction between turn-switch and turn-holding (Hara et al., 2019) or between selecting and non-selecting (Ishimoto et al., 2019) is insufficient to capture the prosodic variation of utterances concerning turn-taking.

5.2. Methods

Selection Type In this analysis, we focus on cases in which substantial utterances reach their possible completion points. The ending types of these utterances include the following three types (see Figure 2): selecting, non-selecting, and multi-unit. Moreover, utterances of the non-selecting type were classified into "turn-switch" (those followed by utterance(s) of the other participant(s), except for response tokens) and "turn-holding" (those followed by utterance of the same participant) depending on the beginning-type tag(s) of the following utterance(s).[2] When the current utterance was followed by utterances of two or more participants and the beginning types of those utterances contained both the turn-switch and the turn-holding types, they were excluded from analysis. Thus, utterances were classified into the following four selection types: selecting, non-sel_switch, non-sel_hold, and multi-unit.

Syntactic Features Part-of-speech (POS) tags provided in the corpus were used as a syntactic feature. The corpus employs the UniDic system for the morphological annotation, which was developed for the morphological annotation of the *Balanced Corpus of Contemporary Written Japanese* (Maekawa et al., 2014). For the current purpose, we merged several POS tags and derived the following five POS types that may be relevant to turn-taking: intj (interjection), pfinal (final particle), pconj (conjunctive particle), vfin (verb and adjective in ending or imperative form), and other (the others).

[2]The beginning-type tags for "turn-switch" are other-selection, self-selection, early-start, and late-start; those for "turn-holding" are continuation and increment.

Prosodic Features Three types of prosodic features were used: the F0, intensity, and average mora duration (AMD) at the final accentual phrase of the utterance. The F0 values were estimated using WaveSurfer, an open-source tool for sound visualization and manipulation, from the speech signals of the final accentual phrases of utterances. The intensity values were calculated in dB for each accentual phrase. The AMDs, related to speech rates in Japanese, were calculated using the time-stamped transcripts provided in the corpus. The F0 and AMD values were log-transformed, and then the three features were each converted into z-scores with respect to individual speakers in order to avoid the influences of sex and individual differences.

Statistical Analysis Statistical analyses were conducted separately for the POS type, F0, intensity, and AMD, all involving the selection type as only independent variable. For the POS type, a mixed-effects multinomial regression model was employed, and for the prosodic features, mixed-effects normal regression models were employed; all models included a random intercept for participants. The estimation of model parameters was performed by using Bayesian inference with a Markov chain Monte Carlo method implemented in the brms package of the R statistical language, which is a wrapper for the probabilistic programming language Stan.[3]

5.3. Results

5.3.1. Syntactic features

Figure 4 shows the marginal effect (the mean and the 95% credible interval) of the selection type at each level of the POS type for the POS-type model, superimposed by the bar plots of the observed POS type frequencies. Pairwise comparisons among the five POS types for each selection type were also performed. Each line in Figure 5 shows the mean and the 95% credible interval of the posterior distribution of the estimated difference between a pair of levels. If the credible interval does not contain the value 0, the difference between the two levels is significant. The results can be summarized as follows:

selecting pfinal was the most frequent, followed by other; intj and pconj were the most infrequent.

pfinal > other > vfin > intj, pconj

non-sel_switch pconj was the most infrequent, followed by vfin; all other POS types were more frequent than these two.

intj, pfinal, other > vfin > pconj

non-sel_hold pconj was the most infrequent, followed by vfin and other; the difference between the frequencies of vfin and other was not significant.

intj, pfinal > vfin, other > pconj

[3]The default settings for the number of Markov chains ($= 4$), the number of total iterations per chain ($= 2000$), and the number of warm-up iterations ($= 1000$) were used. Prior distributions were set to Normal$(0, 10)$ for the selection-type effect and to HalfCauchy$(0, 1)$ for the standard deviations of the random effect and the error term for a normal regression.

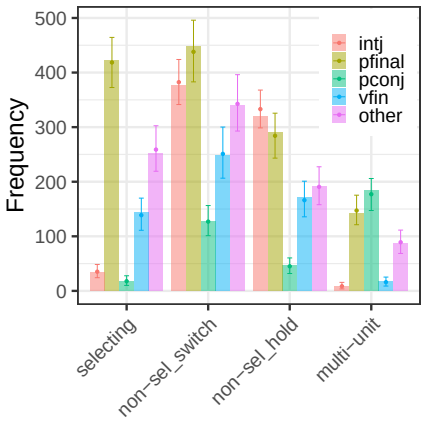

Figure 4: Marginal effect of the selection type at each level of the POS type for the POS-type model

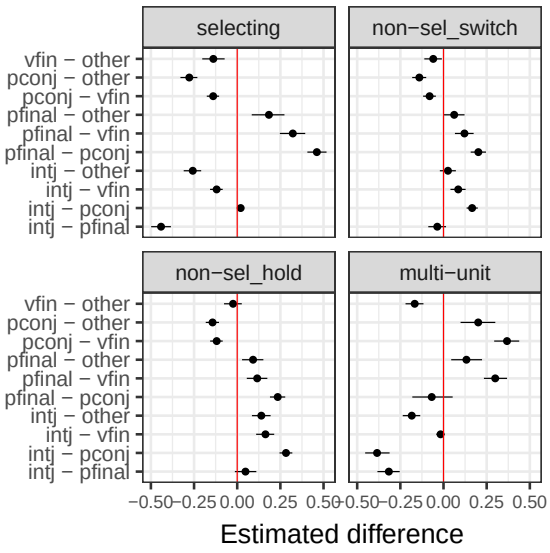

Figure 5: Pairwise comparisons among the POS types relative to the selection type for the POS-type model

multi-unit pfinal and pconj were the most frequent; intj and vfin were the most infrequent;

pfinal, pconj > other > intj, vfin

Overall, the POS-type distributions for non-sel_switch and non-sel_hold were similar to each other, but they were considerably different from the distribution for selecting or that for multi-unit.

5.3.2. Prosodic features

F0 Figure 6 shows the marginal effect of the selection type for the F0 model, superimposed by the violin plots of the observed F0 values. Pairwise comparisons among the four selection types (Figure 7) show that the F0 values for selecting and non-sel_hold were significantly greater than that for non-sel_switch, which was also significantly greater than that for multi-unit. No difference in F0 values, however, was found between selecting and non-sel_hold.

selecting, non-sel_hold >
non-sel_switch > multi-unit

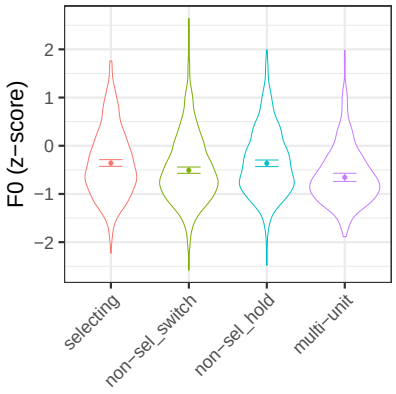

Figure 6: Marginal effect of selection types for the F0 model

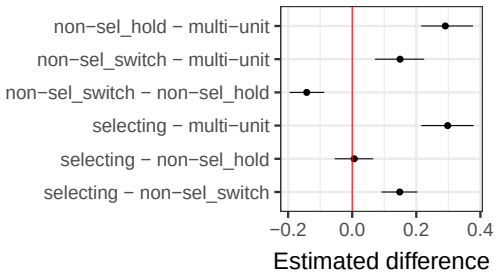

Figure 7: Pairwise comparisons among selection types for the F0 model

Intensity Figure 8 shows the marginal effect of the selection type for the intensity model, superimposed by the violin plots of the observed intensity values. Pairwise comparisons among the four selection types (Figure 9) show that the intensity values for selecting and non-sel_hold were significantly greater than those for non-sel_switch and multi-unit. No difference in intensity values, however, was found between selecting and non-sel_hold or between non-sel_switch and multi-unit.

selecting, non-sel_hold >
non-sel_switch, multi-unit

Average mora duration Figure 10 shows the marginal effect of the selection type for the AMD model, superimposed by the violin plots of the observed AMD values. Pairwise comparisons among the four selection types (Figure 11) show that the AMD values for non-sel_switch were significantly greater than those for the other three selection types. No difference in AMD values was found among these three selection types.

non-sel_switch >
selecting, non-sel_hold, multi-unit

5.4. Discussion

The results clearly show the necessity for a tripartite distinction among selecting (option (a)), non-sel_switch (option (b)), and non-sel_hold (option (c)), suggesting the insufficiency of previous models based on the binary distinction between (a) vs. (b/c) or between (a/b) vs. (c).

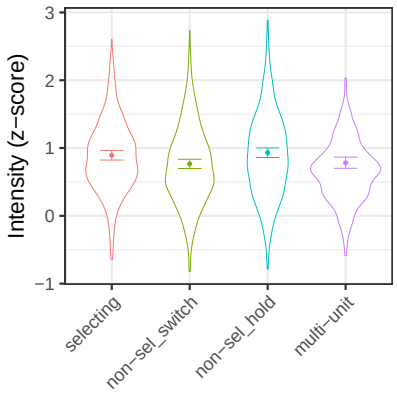

Figure 8: Marginal effect of selection types for the intensity model

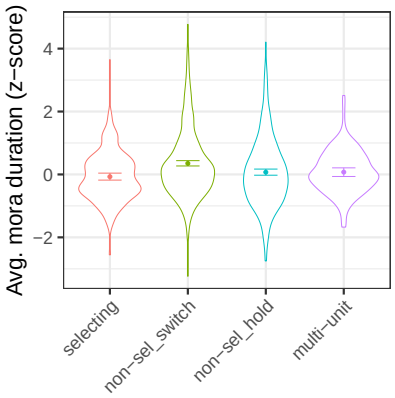

Figure 10: Marginal effect of selection types for the AMD model

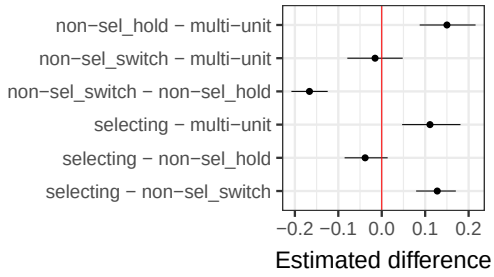

Figure 9: Pairwise comparisons among selection types for the intensity model

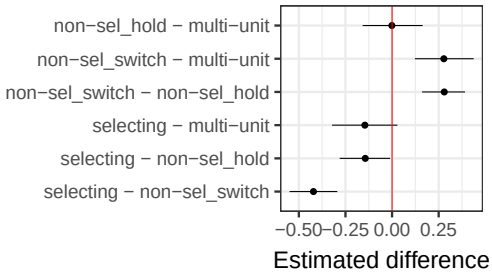

Figure 11: Pairwise comparisons among selection types for the AMD model

If we look at the results more closely, we can make various interesting observations. First, the POS type distributions for `non-sel_switch` and `non-sel_hold` were very similar, but the prosodic features for these selection types were quite different; for all prosodic features, there were significant differences between the two selection types. In syntax both share dominant features displaying no affiliation of a "current-speaker-select-next" technique (`intj` and `other`), but in prosody they exhibit a sharp contrast. Turn-switch at possible completions of the non-selecting type is associated with more decreased F0, weaker intensity, and slower speech rate, while turn-holding at those positions is associated with increased or less decreased F0 and stronger intensity. In short, even when the current speaker has selected no participant as a next speaker, the prosodic features of the utterance display whether the speakership will be shifted to another participant or held by the current speaker.

Second, `selecting` has a unique syntactic property, i.e., the proportion of `pfinal` is much higher than the other POS types, a distribution never observed in the other selection types (Figure 4). In Japanese a first pair part of an adjacency pair, such as a question and request, is typically marked by an utterance-final particle. In Example (5), for instance, *no* (question marker) at the end of the utterance together with a rising intonation indicates that this utterance is designed as a question.[4]

(5) from chiba0232
C: *Osoku nan nai hoo ga ii* ***no?***
late be NEG more NOM good pfinal.Q
'Is (it) better to be not late?'

Since `selecting` is most likely to appear in the first-pair part of an adjacency pair, final particles frequently function as a cue to affiliation of a "current-speaker-selects-next" technique. There are, however, cases where questions are not marked by a final particle, e.g., declarative and echo questions. In most of these, the utterances are prosodically marked by a rising intonation, which is also a property of `selecting` (Figure 6).

Third, `selecting` and `non-sel_hold` share some prosodic features, i.e., increased or less decreased F0 and stronger intensity. Utterances of the `non-sel_hold` type sometimes exhibit an increased F0 in the entire final accentual phrase or at its end.

(6) from chiba0132
C: ***Dotchi demo ii no***
whichever OK pfinal.INT
'Whichever is OK.'
(0.25)
C: *Betsu ni kyoomi ga atte kii teru ka*
meh interest NOM have ask PROG if
doo ka doo demo ii ...
not if doesn't.matter
'It doesn't matter if I'm asking with interest, ...'

[4]In the examples below, the following glosses are used; NOM: nominative case marker, GEN: genitive case marker, Q: question marker, INT: interactive marker, NEG: negative marker, PROG: progressive marker, COP: copula, and N: nominalizer. Numbers in parentheses indicate the duration of a silence (in sec).

In Example (6), C's first utterance is uttered with a relatively high pitch, and it is followed by the same speaker's next utterance after a short pause.

Fourth, multi-unit has also its own syntactic characteristics, i.e., the proportion of pconj is much higher than the other selection types. In Japanese, conjunctive particles like *kedo*, *ga*, and *te* are frequently used to link up successive clauses, resulting in a long stretch of utterance (Iwasaki and Ono, 2001). Thus, the first instance of a conjunctive particle projects "more to come," then the next projects "still more to come," and so on. This is a typical way in which a multi-unit turn is constructed. In Example (7), B's first part of the utterance ends with a conjunctive particle *kedo*, but obviously what he wants to say is not finished yet, thereby projecting "more to come." Indeed, his multi-unit turn is continued after A's acknowledgment.

(7) chiba1232 [simplified]
B: *Baito no hanashi nan*
part.time.job GEN story N
da ***kedo*****[:**
COP pconj.but
'((This is)) a story about my part-time job, but'
A: [*A:, baito ka*
oh part-time.job pfinal.Q
'Oh, ((it's)) part-time job.'
(0.24)
B: *Baito-saki de sa: (0.5) ...*
part.time.job-place at pfinal.INT
'At my part-time job place (0.5) ...'

It is interesting that some prosodic features, i.e., F0 and intensity, for multi-unit were similar to those of non-sel_switch rather than non-sel_hold. Since the status of being inside a multi-unit turn has already been projected by the prior part of that turn, there is no need to indicate it prosodically, in contrast to the case of non-sel_hold. The utterances of the multi-unit type, rather, share some prosodic properties with the utterances of the non-sel_switch type, but the syntactic structure distinguishes the two types.

Our findings are summarized in Table 5. The four selection types can be identified by properties along two dimensions: i) syntactic prominence and ii) prosodic prominence. If the distribution of the POS type for a certain selection type has a distinctively frequent category, such as pfinal for the selecting type and pconj for the multi-unit type, that selection type is "syntacticly prominent"; if no such categories are found, that selection type is "syntacticly mundane." Similarly, if some prosodic features for a certain selection type exhibit prominent values, such as increased or less decreased F0 for the selecting and non-sel_hold types, that selection type is "prosodically prominent"; if no such prominent values are observed, that selection type is "prosodically mundane." Table 5 illustrates how the four selection types are nicely fit in the two-dimensional space defined in terms of the syntactic and prosodic prominences.

Table 5: Syntactic/prosodic prominence and the selection types

	Syntactically prominent	Syntactically mundane
Prosodically prominent	selecting option (a)	non-sel_hold option (c)
Prosodically mundane	multi-unit –	non-sel_switch option (b)

6. Concluding Remarks

In this paper, we proposed a conversation-analytic annotation scheme for turn-taking behavior and demonstrated how it can be used to realize a proper model of turn-taking. In particular, we showed that such a proper model requires the four-way distinction among the types of possible completion of the utterance: i) selecting other participant as next speaker, ii) not selecting next speaker but followed by a switch of the speakership, iii) not selecting next speaker and followed by a continuation of the speakership, and iv) being inside a multi-unit turn.

Although our findings might make a significant contribution to the study of turn-taking models, our study is still preliminary in several points. First, our analysis was only descriptive in that we showed differences of syntactic and prosodic features among selection types but did not construct a model that can predict selection types based on those features. Second, we did not deal with the distinction between completion and non-completion (see Figure 1). Hara et al. (2019) have already shown that the incorporation of this distinction at this level into models improves the accuracy of the prediction. Third, we focused only on the features at or around possible completion of the utterance but did not examine features at other places. Koiso and Den (2010) and Ishimoto et al. (2011) investigated features related to the completion of utterances in Japanese conversations that appear earlier than the final portion of the utterance, and found that some features serve as early cues that project the upcoming completion. Fourth, we did not tackle the task of identifying who is allocated the next turn in the presence of more than one possible next speaker. Several studies (Ishii et al., 2013; Roddy et al., 2018) have incorporated the turn-allocation task into their models using mainly gaze information.

Finally, our annotation could be utilized not only in the study of turn-taking models but also for other areas of conversational studies. For instance, as stated in 3.4., there were atypical combinations of the ending-type tag and the beginning-type tag, such as selecting followed by self-selection or continuation. Investigation into such deviant cases might shed new light on our understanding of sequence organization and conversational structures. All these topics are left for future study.

7. Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number 18K11514.

8. Bibliographical References

Arsikere, H., Shriberg, E., and Ozertem, U. (2015). Enhanced end-of-turn detection for speech to a personal assistant. In *Proceedings of AAAI Spring Symposia*, Mar.

Brusco, P., Pérez, J. M., and Gravano, A. (2017). Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish. In *Proceedings of Interspeech 2017*, pages 2351–2355.

De Ruiter, J. P., Mitterer, H., and Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Den, Y. and Enomoto, M. (2007). A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In T. Nishida, editor, *Conversational informatics: An engineering approach*, pages 307–330. John Wiley & Sons, Hoboken, NJ.

Den, Y., Koiso, H., Maruyama, T., Maekawa, K., Takanashi, K., Enomoto, M., and Yoshida, N. (2010). Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2103–2110, Valletta, Malta.

Den, Y., Yoshida, N., Takanashi, K., and Koiso, H. (2011). Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *Proceedings of the Oriental COCOSDA 2011*, pages 168–173, Hsinchu, Taiwan.

Friedberg, H. (2011). Turn-taking cues in a human tutoring corpus. In *Proceedings of the ACL-HLT 2011 Student Session*, pages 94–98, Portland, OR.

Gravano, A. and Hirschberg, J. (2009). Turn-yielding cues in task-oriented dialogue. In *Proceedings of the 10th Annual Meeting of the SIG on Discourse and Dialogue (SIGDIAL 2009)*, pages 253–261, London, UK.

Gravano, A., Brusco, P., and Štefan Beňuš. (2016). Who do you think will speak next? perception of turn-taking cues in Slovak and Argentine Spanish. In *Proceedings of Interspeech 2016*, pages 1265–1269.

Hara, K., Inoue, K., Takanashi, K., and Kawahara, T. (2019). Turn-taking prediction based on detection of transition relevance place. In *Proceedings of Interspeech 2019*, pages 4170–4174.

Ishii, R., Otsuka, K., Kumano, S., Matsuda, M., and Yamato, J. (2013). Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*, pages 79–86, Sydney.

Ishimoto, Y., Enomoto, M., and Iida, H. (2011). Projectability of transition-relevance places using prosodic features in Japanese spontaneous conversation. In *Proceedings of Interspeech 2011*, pages 2061–2064.

Ishimoto, Y., Teraoka, T., and Enomoto, M. (2019). An investigation of prosodic features related to next speaker selection in spontaneous Japanese conversation. In *Proceedings of the Oriental COCOSDA 2019*, pages 36_1–5.

Iwasaki, S. and Ono, T. (2001). 'Sentence' in spontaneous spoken Japanese discourse. In Joan Bybee et al., editors, *Complex sentences in grammar and discourse*, pages 175–202. Benjamins, Amsterdam.

Koiso, H. and Den, Y. (2010). Towards a precise model of turn-taking for conversation: A quantitative analysis of overlapped utterances. In *Proceedings of DiSS-LPSS Joint Workship 2010*, pages 55–58.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech*, 41:295–321.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48:345–371.

Maier, A., Hough, J., and Schlangen, D. (2017). Towards deep end-of-turn prediction for situated spoken dialogue systems. In *Proceedings of Interspeech 2017*, pages 1676–1680.

Masumura, R., Asami, T., Masataki, H., Ishii, R., and Higashinaka, R. (2017). Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. In *Proceedings of Interspeech 2017*, pages 1661–1665.

Maynard, S. K. (1989). *Japanese conversation: Self-contextualization through structure and interactional management*. Ablex, Norwood, NJ.

Niebuhr, O., Görs, K., and Graupe, E. (2013). Speech reduction, intensity, and F0 shape are cues to turn-taking. In *Proceedings of the 14th Annual Meeting of the SIG on Discourse and Dialogue (SIGDIAL 2013)*, pages 261–269, Metz, France.

Roddy, M., Skantze, G., and Harte, N. (2018). Multimodal continuous turn-taking prediction using multiscale RNNs. In *Proceedings of the 20th ACM on International Conference on Multimodal Interaction (ICMI '18)*, pages 186–190, Boulder, CO.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In Elinor Ochs, et al., editors, *Interaction and grammar*, pages 52–133. Cambridge University Press, New York.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29:1–63.

Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., Taylor, P., Ries, K., Martin, R., and van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41:443–492.

Zellers, M. (2013). Pitch and lengthening as cues to turn transition in Swedish. In *Proceedings of Interspeech 2013*, pages 248–252, Lyon, France.

9. Language Resource References

Yasuharu Den and Mika Enomoto. (2014). *Chiba three-party conversation corpus*. Distributed via NII-SRC.