Annotation of response tokens and their triggering expressions in Japanese multi-party conversations

Yasuharu Den* **Hanae Koiso†** **Katsuya Takanashi‡** **Nao Yoshida§**

*Faculty of Letters, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522, Japan
den@cogsci.L.chiba-u.ac.jp

†National Institute for Japanese Language and Linguistics
10-2 Midoricho, Tachikawa, Tokyo 190-8561, Japan
koiso@ninjal.ac.jp

‡Academic Center for Computing and Media Studies, Kyoto University
Yoshida-hommachi, Sakyo-ku, Kyoto 606-8501, Japan
takanasi@ar.media.kyoto-u.ac.jp

§naou.yoshida@gmail.com

Abstract

In this paper, we propose a new scheme for annotating response tokens (RTs) and their triggering expressions in Japanese multi-party conversations. In the proposed scheme, RTs are first identified and classified according to their forms, and then sub-classified according to their sequential positions in the discourse. To deeply study the contexts in which RTs are used, the scheme also provides procedures for annotating *triggering expressions*, which are considered to *trigger* the listener's production of RTs. RTs are classified according to whether or not there is a particular object or proposition in the speaker's turn for which the listener shows a positive or aligned stance. Triggering expressions are then identified in the speaker's turn; they include surprising facts and other newsworthy things, opinions and assessments, focus of a response to a question or repair initiation, keywords in narratives, and embedded propositions quoted from other's statement or thought, which are to be agreed upon, assessed, or noticed. As an illustrative application of our scheme, we present a preliminary analysis on the distribution of the latency of the listener's response to the triggering expression, showing how it differs according to RT's forms and positions.

Keywords: response tokens, triggering expressions, reaction latency

1. Introduction

Listeners in conversation do not merely hear the speaker's talk, but they sometimes produce small bits of verbal and nonverbal expressions while the speaker's turn is in progress. Among them, verbal response tokens (RTs) have been extensively studied in various fields including discourse analysis, conversation analysis, social psychology, language education, and dialog system research. Japanese RTs, in particular, are known to have plentiful variations in their forms and functions (Clancy et al., 1996; Horiguchi, 1988; Takubo and Kinsui, 1997, *inter alia*). Yet, there have been few studies addressing identification and classification of RTs in real data and attempting to elucidate factors behind the variation of RTs.

Yngve (1970) emphasized the significance of the listener's engagement in an on-going speaking turn, being first to introduce the concept of *backchannels* into the dialog research. Besides typical backchannels like *yes* and *uh-huh*, other types of brief responses that do not claim speakership incipiency, such as acknowledgments and agreements, have been reported in the literature, and methods for automatically discriminating these categories have been developed as part of dialog act modeling (Shriberg et al., 1998, *inter alia*). Few studies, however, have recognized these expressions as a coherent class of listener's responses.

Clancy et al. (1996) studied *reactive tokens*, which cover a wider range of expressions than backchannels, and analyzed their usage in English, Japanese, and Mandarin conversations. They classified reactive tokens into i) backchannels, ii) reactive expressions, iii) collaborative finishes, iv) repetitions, and v) resumptive openers, based on their interactional functions and surface forms. Although they seemed not to provide rigid procedures for annotation, their idea was essential in developing our own scheme.

In conversation analysis, Gardner (2001) compiled RTs reported in the previous studies, classifying them into i) continuers, ii) acknowledgments, iii) change-of-state tokens, iv) assessments, and v) non-verbal responses. He examined usages of eight English RTs, i.e., *mm hm*/*uh huh* (continuers), *yeah*/*mm* (acknowledgments), *oh*/*right* (newsmarkers), and *okay*/*alright* (change-of-activity tokens) from a viewpoint of their interactional functions. An important point of conversation analytic studies of RTs is its emphasis on their roles in sequential organization, i.e., the position in an ongoing sequence, rather than function *per se*.

Following these studies, Den et al. (2011) proposed strict and consistent procedures for Japanese RT annotation, in which RTs are identified and classified according to their forms and sequential positions. Such detailed annotation enables us to investigate a real picture of the variation of Japanese RTs and their correlation with the linguistic and interactional properties such as prosody and sequential context.

Table 1: Form tags

Category	Tag	Example
Responsive interjections	B	*hai*, *un*, *aa*, *ee*, etc.
Expressive interjections	E	*a*, *e*, *hee*, *huun*, etc.
Lexical reactive expressions	L	*soo(-desu-ne)*, *naruhodo*, *tasika-ni*, etc.
Evaluative expressions	A	*sugoi*, *omosiroi-na*, *kowa*, etc.
(Partial) repetitions	R	Repetitions of (a part of) other's speech
(Collaborative) completions	C	One speaker's finishing a prior speaker's utterance

Table 2: Position tags

Category	Tag	Example
First pair parts	1	Request for confirmation or repair of information
Second pair parts	2	Response to a question or request
Sequence-closing thirds	3	Appendix to an adjacency pair
Other responding turns	0	Acknowledgments, assessments, etc.
Unclassifiable positions	9	Signal of self-remembering or self-understanding, marking of topic/activity shift, filling in a break after a topical-talk, etc.
(No position tag		Attention to, understanding of, or evaluation of an on-going turn)

The annotation scheme, however, is still insufficient to deeply study the contexts in which RTs are used, since in this scheme the position of an RT is classified according to its position *in a series of turns*, e.g., the first or second pair part of an adjacency pair (Schegloff and Sacks, 1973), but its position *within the speaker's turn* is not distinguished. Furthermore, it would be an interesting research question how quickly the listener produces an RT upon detecting a responding source in the speaker's turn and how the tendency is different among different forms. To address these issues, this paper proposes a new scheme for annotating *triggering expressions*, which are considered to *trigger* the listener's production of RTs.

In what follows, we first describe the form and the position tags used in our two-stage annotation scheme proposed in Den et al. (2011). We then provide motivation to identify triggering expressions and explain the procedures for annotating them in some detail. We finally present a preliminary analysis on the distribution of the latency of the listener's response to the triggering expression, showing how it differs according to RT's forms and positions.

2. Two-stage annotation of Japanese RTs

In this section, we briefly describe our two-stage annotation scheme of Japanese RTs (Den et al., 2011). In the proposed scheme, RTs are first identified and classified according to their forms, and then sub-classified according to their sequential positions in the discourse.

2.1. Form tags

The following 6 forms are distinguished (Table 1):

1. **Responsive interjections** (B), which express acceptance, at various levels, of an other's utterance, e.g., *hai*, *un*, *aa*, and *ee*, and their successive occurrences.
2. **Expressive interjections** (E), which are used when the listener expresses notice, surprise, disappointment, admiration, etc. elicited by an other's utterance or situation, e.g., *a*, *e*, *hee*, and *huun*.
3. **Lexical reactive expressions** (L), which are short expressions indicating understanding of or agreement with an other's assertion or opinion, e.g., *soo(-desune)* (*I think so*), *naruhodo* (*really*), and *tasika-ni* (*surely*).
4. **Evaluative expressions** (A), which assess the talk of the prior speaker, usually realized by short adjectives or adjective verbs such as *sugoi* (*great*), *omosiroi-na* (*funny*), and *kowa* (*terrible*).
5. **(Partial) repetitions** of other's speech (R), which are sometimes used to express an understanding of or agreement with the information conveyed by an other speaker.
6. **(Collaborative) completions** (C), where one speaker finishes a prior speaker's utterance, predicting what would follow the part of the utterance produced so far.

2.2. Position tags

The position tag is intended to capture substantial functions that RTs may serve beyond simply signaling listener's attention and involvement; these functions include an affirmative answer to a question, an acceptance of a request, a repair initiation when affiliated with an interrogative intonation, and so on. The following 5 positions are distinguished (Table 2):

1. **First pair parts** of adjacency pairs (1), where RTs are used, typically accompanied by an interrogative intonation, to elicit an addressee's response such as confirmation or repair of information.
2. **Second pair parts** of adjacency pairs (2), where RTs are used to respond to an other's elicitation such as a question or a request.

3. **Sequence-closing thirds** (3), which are sometimes appended to an adjacency pair, designed to move for sequence closing (Schegloff, 2007), typically realized by a brief item like *aa* or *un* as well as an assessment.

4. **Other responding turns** (0), which are other positions than the above three and in which RTs occupy a full turn, or a preface to it, not just inserted as a recipient's reaction but committed to some degree of speakership incipiency; typical examples are acknowledgments and assessments.

5. **Unclassifiable positions** (9), which are other cases where tokens in the form of an RT appear to occupy a full turn; they are used to signal self-remembering or self-understanding, mark topic/activity shift, fill in a break after a topical-talk, and so on.

The first two positions are based on the concept of adjacency pairs (Schegloff and Sacks, 1973) and the third one on its extension (Schegloff, 2007). The fourth position, 0, however, is different from the second one, 2, in that position 2 is prospectively occasioned, its absence being noticed as such, while position 0 is connected to the previous utterance retrospectively.[1]
RTs that do not appear at the above 5 positions are left without being assigned a position tag. They occur at "with-in-turn" position, and typically indicate attention to, understanding of, or evaluation of an on-going turn.

3. Annotation of triggering expressions

Although the RT annotation scheme described in the previous section provides strict and consistent procedures, which are prerequisite to ensure reliability and reproducibility, it is still insufficient to capture the contexts in which RTs are used, because it lacks information about the source of the listener's response. To overcome the weakness, we extend our scheme to include procedures for identifying expressions in the speaker's turn that *trigger* the listener's production of RTs.

3.1. Motivation

It has been shown that RTs associated with different functions often appear at different positions in relation to the prior turn. For instance, continuers, which do not claim speakership incipiency but merely signal a 'go-ahead' sign to the speaker, are typically located at boundaries of utterances (Schegloff, 1982), while listeners may produce an assessment on a particular object or proposition within an on-going turn (Goodwin, 1986). These findings suggest that in order to understand the function of RTs, we have to identify not only their positions in the conversational sequence but also their positions within the speaker's turn. For this purpose, we propose a new scheme to annotate *triggering expressions*.

[1] In some traditions of discourse analysis (Coulthard, 1985), position 0 is treated as 'second' in a similar sense as the second pair part of an adjacency pair.

3.2. Types of triggering

In this new scheme, RTs, recognized by the above two-stage annotation scheme, are further classified according to whether or not there is a particular object or proposition in the speaker's turn for which the listener shows a positive or aligned stance. Either of the following categories is assigned.

1. **Object** or **proposition**. These are used when some expression in the speaker's turn can be seen as triggering such reaction of the listener as agreement, assessment, and notice of a newsworthy thing. They are distinguished according to whether such expression represents an object or a proposition.

2. **No-trigger**. This is used when there is no particular expression in the speaker's turn that may trigger the listener's production of the RT. In these cases, the listener's reaction may show a signal of continued attention, understanding of the meaning being conveyed, or passing up an opportunity to take a turn (Schegloff, 1982), or deliver an affirmative answer to a question or request, rather than a positive or aligned stance to the speaker. The category also includes cases where RTs are produced in a self-motivated way, rather than evoked by an other's utterance.

3.3. Triggering expressions

For RTs of type 'object' or 'proposition,' the triggering expression is identified in the speaker's turn. Triggering expressions include

- surprising facts and other newsworthy things,
- opinions and assessments,
- focus of a response to a question or repair initiation,
- keywords in narratives, and
- embedded propositions quoted from other's statement or thought,

which are to be agreed upon, assessed, or noticed. For simplicity, we annotate only the rightmost words of triggering expressions, i.e., head nouns for objects and verbal components for propositions, which, in Japanese, are both placed on the right-edge of a phrase or clause.

3.4. Examples

Figure 1 shows some examples of our RT annotation, i.e., form and position tags and triggering expressions. RTs are shown in boldface, with the form tag indicated by a capital letter after '/,' possibly followed by the position tag written with a digit; their types of triggering are also given after '=.' The (rightmost words of the) triggering expressions are shown in double angle brackets and co-indexed with the RTs that are triggered by these expressions.
For instance, in (1), B's *aa* is an expressive interjection, which occurs within C's on-going turn, hence no position tag; immediately following is B's another RT, *Huziya-hoteru* (the name of a famous hotel), which is a repetition

(1) chiba0632:412.11-414.44
C:Huziya-⟪hoteru⟫$_1$-ga [hurin-ga ⟪ooi⟫$_2$-t[te sensee-ga yu-
At Fujiya Hotel, many guests have affairs, our professor sai-
B: [**aa/E=object**$_1$;notice of
Huziya-hoteru/R=object$_1$;a newsworthy thing
Ah, Fujiya Hotel.
A: [**a/E=proposition**$_2$;agreement
soo/L=proposition$_2$;with a quote
Oh, right.

(2) chiba1032:175.57-176.81
B:o-nee-san ⟪kiree⟫$_3$-da-[yo-ne
The lady is beautiful, isn't she?
C: [⟨laugh⟩
A:**un/B2=proposition**$_3$;alignment with a request for agreement
Yes, I agree.

(3) chiba0932:512.30-518.74
B:de doitu-go-wa yome:-te hatuon-wa dekiru-kedo[:
And, as for German, I can read and pronounce it, but
C: [**un/B=no-trigger** ;continuer
uh-huh
B:syabe-re-nai-ne
I can't speak it.
(1.4)
B:**soo/L9=no-trigger** ;self-understanding
Yes.

(4) chiba0632:548.40-550.67
B:Negisi-san-mo yoku it-te-rassyai-masu-yo-ne
Ms Negisi, I think you, too, often go to
B:Hakone
Hakone.
C:**soo/L2=no-trigger** ;affirmative answer that is sequentially elicited
Yes.

Figure 1: Examples of RT annotation

of the initial part of C's prior utterance. Since *aa* is a typical change-of-state token (Heritage, 1984) in Japanese, this *aa*-prefaced repetition can be regarded as indicating the listener's notice of a newsworthy thing, which is relevant to *Huziya-hoteru*. Thus, these two RTs are labeled 'object,' and their triggering expression is identified as *Huziya-hoteru* in C's utterance, whose rightmost word *hoteru* is bracketed in example (1).

Similarly, A's *a* and *soo* are an expressive interjection and a lexical reactive expression, respectively, both with no position tag, which are considered to show agreement with a quoted statement in C's utterance, *Huziya-hoteru-ga hurin-ga ooi* (*At Fujiya Hotel, many guests have affairs*), uttered by their professor. Thus, these two RTs are labeled 'proposition,' and their triggering expression is identified as the quote in C's utterance, whose rightmost word *ooi* is bracketed in example (1).

On the other hand, in (3), C's *soo* is produced after a lapse of 1.4 seconds, which is preceded by her own utterance, that has reached a possible completion point of her turn. This RT can be seen as signaling self-understanding or filling in a break after a topical-talk; thus, its position tag is `9` and its type of triggering is 'no-trigger.' In (4), C's *soo* is an affirmative answer to B's request for confirmation, *Negisi-san-mo yoku it-te-rassyai-masu-yo-ne, Hakone* (*Ms Negisi* (= C's name)*, I think you often go to Hakone, too*); thus, its type of triggering is also 'no-trigger,' although its status as an affirmative answer is represented by position tag `2`.
There is no straightforward correspondence between types of triggering and RT forms/positions. For instance, the same token *soo* has a triggering expression in (1) but no triggering expression in (3) or (4); the same applies to *un*s in (2) and (3). Similarly, tokens at the same position, e.g., the second pair part of an adjacency pair, may have a triggering expression or may not as in (2) and (4).

4. Data and analysis

In this section, we apply our annotation scheme to our multi-party conversation data, and present a preliminary

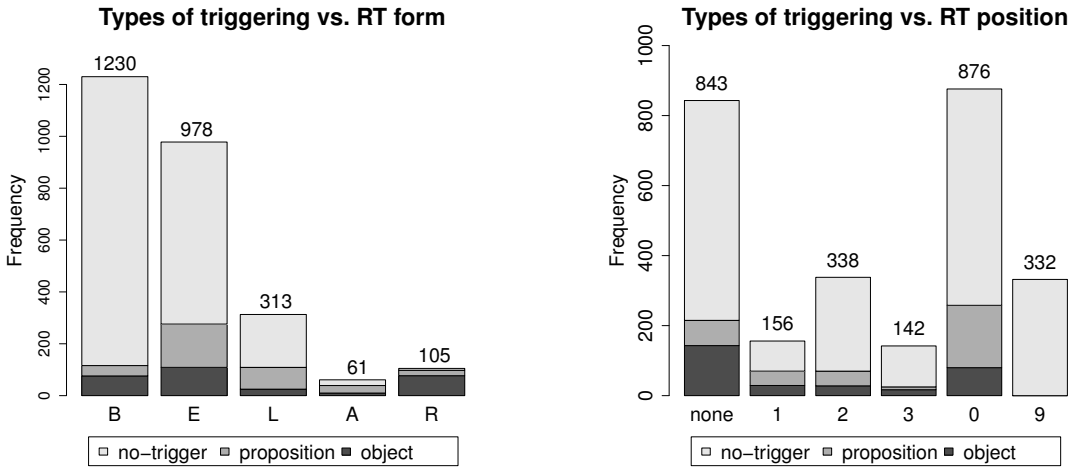

Figure 2: Distributions of types of triggering in relation to RT forms and positions (the left and the right graphs, respectively). The number on each bar indicates the total number of the cases in that RT form/position category.

analysis on the distribution of the latency of the listener's response to the triggering expression, showing how it differs according to RT's forms and positions.

4.1. Data

Twelve dialogs, produced by 36 speakers, were selected from the *Chiba three-party conversation corpus* (Den and Enomoto, 2007). The corpus is a collection of casual conversations among three participants. The participants of each dialog were friends on campus. Each dialog is about 10 minutes long, and a total of 2 hours of dialogs were used in this analysis. The corpus has already been annotated with various sorts of information such as utterance-units, morphological information, prosodic information, and RT forms and positions (Den et al., 2010; Den et al., 2011).

Two of the authors conducted annotations of RT forms and positions and triggering expressions. They first worked on 2 dialogs independently and discussed for inconsistency: the agreement was $\kappa = .70$ and $\kappa = .50$. After reaching tentative consensus on detailed criteria, each of these authors labeled each half of the remaining 10 dialogs.

A total of 2693 RTs were identified in the data. Among RTs of the 6 forms, completions were excluded from the analysis due to a small number of cases ($N = 6$).

4.2. An illustrative analysis

Figure 2 shows the distributions of the types of triggering, i.e., 'object,' 'proposition,' and 'no-trigger,' in relation to the 5 RT forms, excluding completions, and the 6 RT positions, including 'no position tag' category. Obviously, the distribution varies depending on forms and positions. RTs of form other than responsive interjection (B) are more often triggered; in particular, evaluative expressions (A) and repetitions (R) involve triggering expressions 64% and 93% of the time, respectively. RTs appearing at the first-pair-part position (1) are more often triggered as well, i.e., 45% of the time, compared to the other positions, at which RTs have triggering expressions only 18–29% of the time.[2]

[2]By definition, RTs appearing at an unclassifiable position (9) cannot have triggering expressions.

Figure 3 shows the distributions of reaction latency of RTs, measured from the end of the triggering expression and the utterance-unit containing it, in relation to the types of triggering ('object' vs. 'proposition') and the 5 RT forms. For object-type triggering expressions, the distribution of the reaction latency exhibits marked differences among forms. Responsive interjections (B) and lexical reactive expressions (L) are likely to be produced more quickly upon detecting the triggering expression than the other RTs; the medians of reaction latency in B- and L-form RTs are smaller than the others, when measured from the end of the triggering expression (see the leftmost graph). On the other hand, it is evident that the production of evaluative expressions (A) and repetitions (R) are postponed until the speaker's utterance reaches its completion; the medians of their latency are close to or greater than zero, in contrast to those of the other RTs, whose medians are less than zero, when measured from the end of the utterance-unit containing the triggering expression (see the second graph).

For proposition-type triggering expressions, responsive interjections (B) and lexical reactive expressions (L) are likely to be produced more quickly upon detecting the triggering expression than the other RTs, a similar tendency as was observed for object-type triggering expressions (see the third graph). The tendency that evaluative expressions (A) and repetitions (R) are likely to be produced around the end of an utterance-unit is also replicated; however, it is not peculiar to these forms of RTs but it may be common to other forms (see the rightmost graph); the medians of reaction latency in responsive interjections (B) and expressive interjections (E) are also close to or greater than zero. This may be due to the difference of the location in which object-type and proposition-type triggering expressions reside in the speaker's utterance.

In sum, the distributions of types of triggering and reaction latency showed considerable differences depending on forms and positions of RTs. These results encourage us to utilize our annotation scheme for investigating the relationship between triggering expressions and the variation of Japanese RTs in the future research.

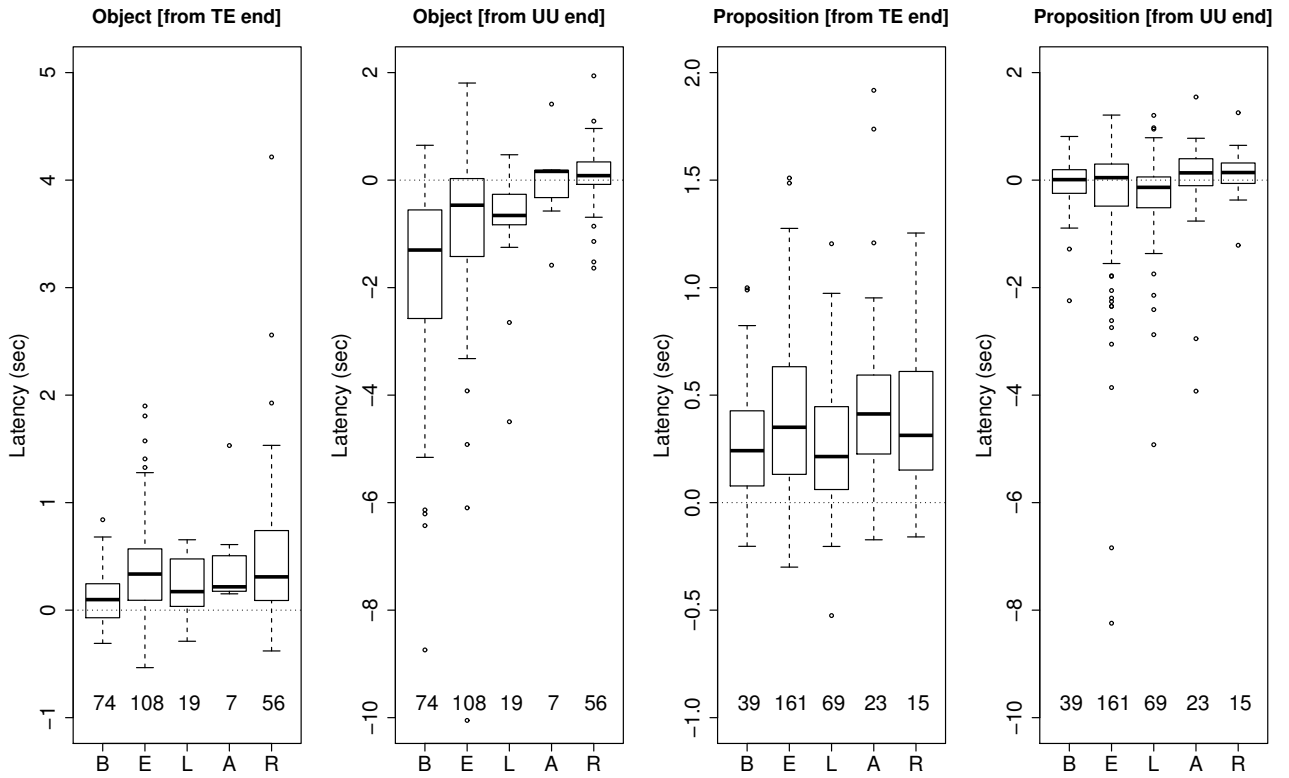

Figure 3: Distributions of reaction latency of RTs, measured from the end of the triggering expression (TE) and the utterance-unit (UU) containing it, in relation to types of triggering and RT forms. The number below each boxplot indicates the total number of the cases in that RT form category. Some extreme outliers are not shown.

5. References

P. M. Clancy, S. A. Thompson, R. Suzuki, and H. Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26:355–387.

R. M. Coulthard. 1985. *An introduction to discourse analysis*. Longman, London, 2nd edition.

Y. Den and M. Enomoto. 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.

Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida. 2010. Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010)*, pages 2103–2110, Valletta, Malta.

Y. Den, N. Yoshida, K. Takanashi, and H. Koiso. 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.

R. Gardner. 2001. *When listeners talk*. John Benjamins, Amsterdam.

C. Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*, 9:205–217.

J. Heritage. 1984. A change-of-state token and aspects of its sequential placement. In J. M. Atkinson and J. Heritage, editors, *Structures of social action: Studies in conversation analysis*, pages 299–345. Cambridge University Press, Cambridge.

S. Horiguchi. 1988. Hearers' behaviors in communication (in Japanese). *Journal of Japanese Language Teaching*, 64:13–25.

E. A. Schegloff and H. Sacks. 1973. Opening up closings. *Semiotica*, 8:289–327.

E. A. Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen, editor, *Analyzing discourse: Text and talk*, pages 71–93. Georgetown University Press, Washington, D.C.

E. A. Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis 1*. Cambridge University Press, Cambridge.

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

Y. Takubo and S. Kinsui. 1997. The discourse management function of fillers in Japanese (in Japanese). In Spoken Language Working Group, editor, *Speech and grammar*, pages 257–279. Kuroshio Shuppan, Tokyo.

V. Yngve. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–578, Chicago.