

# Joint Construction of Narrative Space: Coordination of gesture and sequence in Japanese three-party conversation

Eiji Toyama<sup>1</sup>, Kouhei Kikuchi<sup>2</sup>, and Mayumi Bono<sup>2</sup>

<sup>1</sup> Advanced Integration Science, Chiba University, 1-33, Yayoi-cho, Inage-ku, Chiba, Japan

<sup>2</sup> National Institute of Informatics, 2-1-2, Chiyoda-ku, Tokyo, Japan

<sup>1</sup>etohtyama@cogsci.L.chiba-u.ac.jp, <sup>2</sup>{vadise, bono}@nii.ac.jp

**Abstract.** We investigate how participants establish co-tellership by coordinating utterances and gestures in Japanese three-party conversation. When speakers in Japanese conversation tell shared knowledge (co-telling conversation), they tend to collaboratively build sentences. Hayashi[1] describes how Japanese establish co-tellership not only by using utterances but also by coordinating their body movements and gazes. But Hayashi[1] did not describe how people use gestural expressions in the co-telling conversation. To investigate this, we especially focus on participants' gestural expressions and their gestural viewpoint. In this paper, we describe the following findings: 1) participants who collaboratively tell a story (co-tellers) tend to use similar gestural expressions. 2) co-tellers do not only share their gestural expressions but also share their gestural viewpoint. 3) co-tellers do not only coordinate their own speech and gestural expression but also aid in the coordination of each other's multimodal expression.

**Keywords:** .Conversation analysis, gesture, joint construction, co-tellership, repair

## 1 Introduction

### 1.1 Co-tellership in Japanese

Speakers in Japanese conversation tend to tell shared knowledge collaboratively. This collaborative storytelling is referred to as co-tellership. In a co-telling conversation, people typically adjust their role from moment to moment. Hayashi[1] describes how people establish co-tellership not only by using utterances but also by coordinating their body movements and gazes.

### 1.2 Viewpoint in speech and gesture

McNeill points out that two different viewpoints are expressed through the kinds of gestures that speakers use during narratives[2]. He writes “Basically, two viewpoints

appear in the gestures people perform during narratives: the gesture may seem to reenact the character, and this is character viewpoint (C-VPT), or the gesture may appear to display an event from the viewpoint of an observer, and this is observer viewpoint (O-VPT).” McNeill also points out that the character viewpoint gestures tend to appear when the speaker narrates a central content in a story[2]. However McNeill focused on not co-telling conversation but monologue in narrative talk. It is unclear whether McNeill’s theory is applicable to co-telling conversations. Joh et al.[3] investigated how people coordinate simultaneous gestural matching in which the participants tell the story collaboratively. Joh et al. revealed how co-tellers cooperate to produce the same gesture at the same time. However, Joh et al.[3] did not mention how the gestural viewpoints are shifted between the co-tellers in sequence.

The gestural viewpoint is very important because it determines the layout of the gesture, in other words, it determines how the gesture is expressed. We investigate how the gestural viewpoint is shared between co-tellers.

### **1.3 Research objectives**

The purpose of this study was to investigate how speakers establish co-tellership by using gestural expressions in conversation. In particular, this paper focuses on the following three aspects: 1) how the participant coordinates co-tellership in conversations not only by producing utterances but also by adding accompanying gestural expressions,; 2) how they share viewpoint for producing speech and gestures,; and 3) how they organize the sequential structure, especially repair sequence, of co-telling conversations.

## **2 Data**

### **2.1 Data collection**

The following transcript is part of a three-party conversation that was videotaped by seven cameras in IMADE room at Kyoto University on July 2010. The IMADE (Interaction Measurement, Analysis, and Design Environment) room was developed as an environment for recording human conversational interactions[4]. This environment has a human motion capture system and a gaze tracking system. The participants did not know each other before the experiment. Two of the participants, S2 and S3, were instructed to watch the cartoon film, “Canary Row” in another room, and subsequently tell the story of the film to the other participant, S1. Each of participants put on the markers and gaze tracking devices for capturing their movement and gaze direction. In this paper, however, we analyze video and hand annotation data for gestural movement.

## 2.2 Transcription and annotation

The following conversation data is composed by Japanese transcript (Roman style) and English translation of a scene in the data. Japanese transcript is marked by [ ] that means overlapped with interlocuter's utterances. The listening participant, S1, is at the lower-right in the Fig.1. S2, one of the co-tellers, is at the upper-right in the Fig.1, and S3 is on the left side. The transcript consists of three layers: the numbering layer is the main layer of this transcript. This layer describes utterances in romanized Japanese. The second layer is the gloss of each Japanese word. The third layer is a translation of Japanese into English.



Fig.1: Sitting Position

### Japanese transcript (Roman style) and English translation

01 S2: *madogiwa ni Tweety no torika[go ga oitea(0.4)]tte*  
 "Tweety's bird cage is put by the window."  
 02 S3: *[un-ah[ha*  
 "Uh huh-Oh h"  
 03 S1: *[un*  
 "Uh huh"  
 04 S2: *de neko ga sore wo mitsukete (0.6) [nera(0.8)]tte*  
 "Then, the cat notices it and sets his sights on  
 it(Tweety)."  
 05 S3: *[saisho boenkyo de*  
 06 *(0.41) tweety wo miterundesuyo*  
 "At first, (the cat) is watching Tweety with a  
 binocular."  
 07 *torikago ni haitteru (0.6) nerattete soshitara (0.23)*  
 "(The cat) is focusing on (Tweety) in the birdcage. Then,"  
 08 *tweety mo (0.68) boenkyo de mite[(mashite) [ hhh*  
 "Tweety is also watching the cat with a binocular."  
 09 S2: *[sou sou*  
 "yeah yeah"  
 10 S1: *[n hhh*  
 11 S2: *mite[te*  
 "(Tweety) is watching."  
 12 S1: *[otagai mi[atterundesuka*  
 "Are they watching each other?"  
 13 S3: *[tagai miatte de Tweety ga*

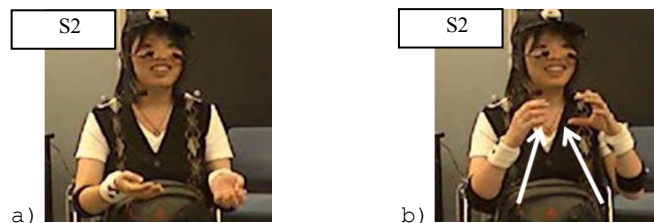
14        *nerawareteru (1.17)*  
 15        *te waka(0.44)tte*  
**"They are watching each other. And Tweety sees that he**  
**is being targeted"**  
 16        *(0.92) a [chigauka*  
               **"oh no"**  
 17 S2:        *[de (0.4) neko:: [ha neratte: de (0.36)*  
**"Then, the cat is focusing on (Tweety). Then,"**  
 18 S3:        *[un*  
                   **"yeah"**  
 19        *hoteru: no shoumen kara (0.49)*  
 20        *ma: haitteikunda kedo(0.3)*  
**"(The cat) goes in the hotel's main entrance, but**  
 21        *oikaesareru to*  
**(The cat) is made to go away."**

### 3 Coordination of gesture


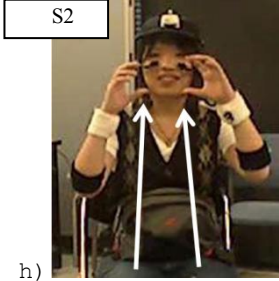
In this section, we describe in detail how S2 and S3 achieved co-tellership through their gestures. Particularly, we focus on the gestures which imitated the previous speaker's gesture. We term these gestures "mirroring gestures". We explain that all of the mirroring gestures occurred when the main teller switched between S2 and S3.

The following segments are parts of the Japanese transcript. When the utterance accompanies gestural expressions, the gesture phase is described as the first layer. The gesture phase consists of four phases[5]: 1) "prep" means preparation for the stroke movement. 2) "str" means stroke, which is the main movement of the gesture. 3) "hld" means to hold the movement before or after the stroke. 4) "ret" means retraction, which is the movement to get back to the home position. The other exceptional phases are following; "X.stop", "RP", "retRP". "X.stop" means that some gestural phase are aborted, where X is variable. "RP" means rest position in which the speaker's hand is raised but relaxed after gestural movements. "retRP" means retraction movement to RP. A gesture unit is composed by one or several gesture phases; preparation, stroke, and retraction[5]. The gesture unit is marked by { }. The pictures below of the gesture phase show the start and end of the stroke.

#### Segment (1)





		↓ e)		↓ f)
	/RP	/prep	/str/hld	
08 S3	tweety mo	(0.68) boenkyo	de mite[(mashite) [ hhh	
	Tweety	binocular	watch for	
	<b>"Tweety is also watching the cat with a binocular."</b>			
		↓ g)	↓ h)	
		{/prep	/str/ret	
09 S2:		[sou sou		
		<b>"yeah yeah"</b>		
	S2	S2		
				
	g)	h)		

In line 09, S2's movement from g) to h) occurred at the time when S2 was saying, "sou sou (yeah yeah)" and when S3 was making the binocular gesture. According to picture e), S3 was still making the binoculars gesture and gazing at S2. In the cartoon film, when the cat found Tweety, Tweety was also watching the cat with his binoculars. By making the binoculars gesture, S3 seemed to replay the scene from Tweety's viewpoint.

Thus, as mentioned above, the mirroring gestures occurred when the main teller switched between S2 and S3. In study of individual gestural expression, it is said that repetition of similar gestures makes cohesion of discourse. This repetition of gestures is termed catchment[7]. In the point of repetition of gestures in co-telling conversation, the mirroring gesture may have the same function.

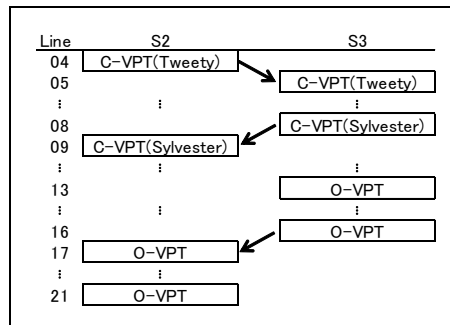
#### 4 Sharing viewpoint

In this section, we review the above gestures, and focus on the viewpoint of producing gestures. As we see in the following, S2 and S3 not only imitated each other's gestures but also adopted the gestural viewpoint.

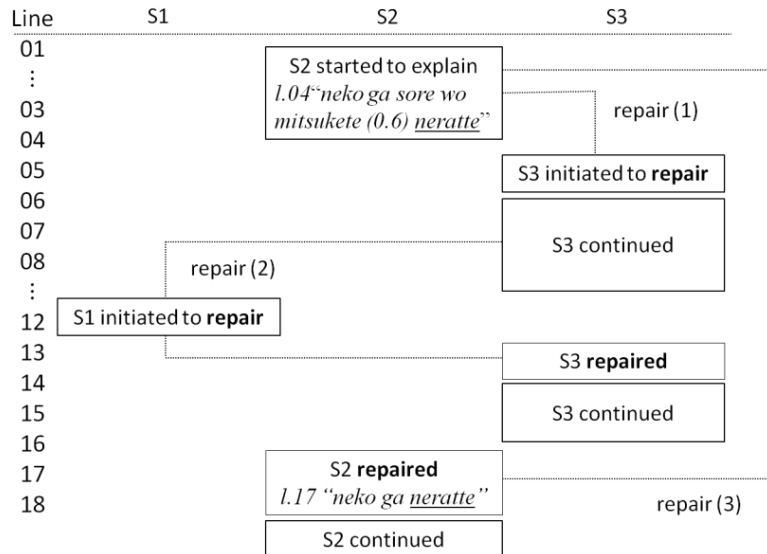
The binoculars gestures, a)-b) and e)-f) in the example, are character viewpoint gestures. More properly, the c)-d) movement and the e)-f) movement are displayed from Sylvester's viewpoint. The g)-h) movement is displayed from Tweety's viewpoint. Thus, these gestures resemble each other in movement but represent different character viewpoints.



S3 continued to explain. However, she finally aborted the explanation at line 16. At the same time, S2 started to explain this part of the story with an accompanying gestural expression. This gesture was produced from the same observer viewpoint that S3 used in the previous line. Then, S2 finished explaining the rest of the episode. Note that S2 did not only imitate S3's character's perspective (Sylvester or Tweety) but also took over the framework of the explanation (Character or Observer). In short, the achievement of co-tellership is understandable from their gestural expression and viewpoint.



**Fig.2: Transition of viewpoints of gestures**



**Fig.3: Three repairs in the data**

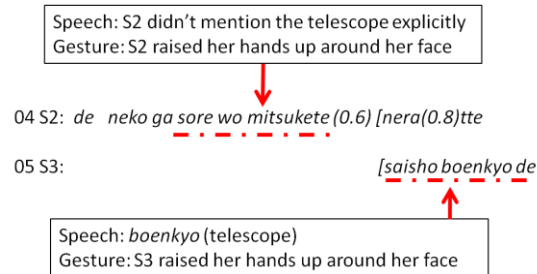


## 5 Repair in co-telling conversation

There are three repairs in this data (Fig.3): 1) in lines 04-05 S3 repaired S2's mismatch of multimodal expression. 2) In lines 08-13 S1 initiated to repair what S3 had said and then S3 restated the scene in a different viewpoint. 3) In lines 16-17 S2 repaired again what S2 could not completely repair in line 04. This self-re-repair suggests that S2 argued S2 had epistemic priority. In the following sections, we describe these repairs in detail.

### 5.1 Repair (1): Repairing combination of speech and gesture

Segment (1) is not only an example of mirroring gesture but also an example of a combination of speech and gesture repair. In lines 04-05, S3 repaired what S2 described in gestural expression but not in speech. In line 04, S2 did not explicitly mention the binoculars which Sylvester used but raised her hands to her face (see segment (1) and Fig.4) to signify "binoculars". This mis-matched multimodal expression became a source of miscommunication that S2 would attempt to repair. S2 actually tried to repair her mismatch expression after the pause(0.6 sec) in line 04, however S3 interrupted S2 and began to repair S2's mistake by saying, "*saisho boenkyo de*". Thus the repair (1) is a case of *other-initiated other-repair*. Note that we could not understand why S2 fell silent and why S3 could take the next turn if we did not first focus on both S2's speech and gestural expression.



**Fig.4: Combination of speech and gesture in Segment (1)**

**Table.1: Sequential relation of verb type and gestural viewpoint**

Line	Speaker	Subject	Verb type	Gestural viewpoint
05	S3	Sylvester	uni-directional	C-VPT
:	:	:	:	:
08	S3	Tweety	uni-directional	C-VPT
:	:	:	:	:
12	S1	Sylvester and Tweety	bi-directional	N/A
13	S3	Sylvester and Tweety	bi-directional	O-VPT

## 5.2 Repair (2): Repairing viewpoint

According to Fig.2 and Table 1, S3's gestural viewpoint changed at line 13. In this section we describe why S3 changed the gestural viewpoint from C-VPT to O-VPT. At line 12, S1, the listener, asked S3 "*otagai miatterundesuka* (Are they watching each other?)". In other words, S1 initiated to repair what and/or how S3 had said before line 12. S1's question is important because using the bi-directional verb "*miatteru* (watch each other)" prompted S3 to repair the following two points of S3's explanation before line 12.

S3 initially described Sylvester and Tweety in sequence; Sylvester was the subject from line 05 to line 07 and Tweety was the subject in line 08. The important point of this scene is that Sylvester and Tweety are watching each other at the same time. However this simultaneity was lacking in S3's spoken explanation. With regard to gestural expression, S3 adopted C-VPT from line 05 to line 08. Thus, in these lines, S3 only mentioned one character at a time both in speech and in gestural expression. Moreover, S3 used uni-directional verbs from line 05 to 08; "*miteru* (watch)", "*nerattete* (aim for)", "*mite* (watch)". These uni-directional verbs do not express the simultaneity of Tweety and Sylvester's mutual gaze. These two problems, the sequential description of Tweety and Sylvester, and the uni-directional verbs that S3 used, explains why S1 would be confused.

In line 12, S1 presented her understanding of the scene by asking, "*otagai miatterundesuka* (are they watching each other?)". Then in line 13 S3 repaired her expression using the bi-directional verb and O-VPT gesture (Table.1). The repair (2) is a case of an *other-initiated self-repair*. In this way, S3 could describe two characters at a time. Thus, the target of repair is not only speech but also the expression of gestural viewpoint.

## 5.3 Repair (3): Sequential design regarding epistemic priority

In assessment sequence, it is said that there are some asymmetries between a person who tells his/her assessment first and the follower because a first assessment may be regarded as claiming epistemic priority[8]. In this data, there is no assessment sequence but there may be an epistemic priority between S2 and S3 because there is a main teller. In this section, we describe how S2 argued her epistemic priority by analyzing her speech and the participants' gazes. We added three gaze layers to segment (4), see segment (5) below. Each of the gaze layers has the following components. "S1", "S2" or "S3" means that the gaze direction was fixed to that participant. "t" means transition from a participant to another. "up" means gaze direction was not fixed to someone and was fixed to upward. "S3.hand" means that S3 was watching to her own hands.

Repair (3) is an example of *re-repair*. In line 04, because of the mismatched multimodal expression, S2 could not explain the scene enough. After the 0.6 sec pause, S2 tried to repair her explanation, saying "*nera(0.8)tte* (focusing on)". But this *self-initiated self-repair* was not completed because S3 interrupted with her own repair on line 05. Comparing line 04 with line 17, we can find the same expression,

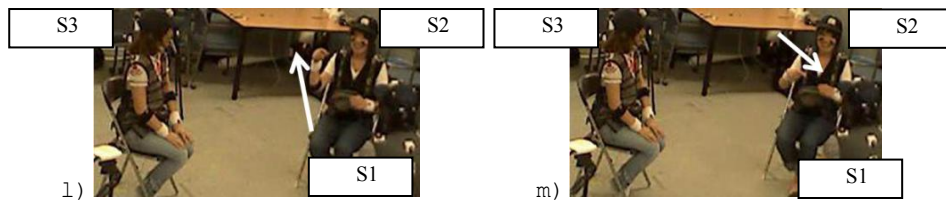
“neratte”, which was used to try to repair in line 04. In line 17, after adjusting narration environment, S2 used the same expression again to repair, countinuing where she left off before she was interrupted. Now we describe how S2 adjusted the narration environment below.

Before index 1) S1 and S2 were looking at S3 and then S2 was looking at S1. On index 2) S3 was looking at S2 at the same time S3 stopped telling the story. On index 3) S1 saw S3 starting to look at S2 and S1 also began to see S2. Then S2 started to watch to S3 so that S2 and S3 were watching each other at the time S3 gave S2 back-channel feedback in line 18. On index 4), after receiving the back-channel feedback, S2 was looking at S1, saying “neratte” with gestural expression.

From index 1) to 3), S2 would confirm the narration environment to tell about the episode. S2 began to speak on index 2) but she would wait until she had the other participants’ gaze by using the 0.4 sec pause. After getting the back-channel feedback, on index 4), S2 initiated to re-repair with a gesture that was not accompanied in line 04. Thus, adjusting the narration environment, S2 used the same expression “neratte” effectively to repair. This re-repair suggests that the sequence from line 05 to 16 was a sub-sequence of S2’s narrative sequence. In that sence, using the re-repair, S2 may have argued that she is the main teller.

### Segment (5)

	↓1)	↓2)	↓3)	↓4)	
S1-gaze:	S3-----	-----	/t/S2-----	-----	
S2-gaze:	S3--/t/S1-	-----	-/t--/S3--/t/S1----	-/t/up-----	
S3-gaze:	S3.hand---	/t---/S2-----	-----	-/t---/S1-----	
16 S3:	(0.92)	a	[chigauka		
		"oh no"			
17 S2:		[de (0.4) neko::	[ha neratte:	de (0.36)	
		"Then, the cat is watching. Then,"			
18 S3:		[un			
		"yeah"			



## 6 Conclusion

When people collaborate to tell a story, they complement each other’s explanation in the phenomenon known as co-tellership. Our research shows evidence that they also coordinate gestural representation to aid the co-tellership.

The co-tellership in this study consists of three components. First, as described in section 3, the mirroring gestures appear when the next speaker starts the turn by producing utterances. Second, as described in section 4, the co-tellers do not only imitate each other's gestural expressions but also take over the framework of the explanation. This suggests that the co-tellers share the knowledge about the story in detail. Third, as described in section 5, S3 began to repair S2's mismatched multimodal expression (repair 1) and S1 began to repair S3's gestural viewpoint (repair 2). This suggests that they do not only coordinate their own speech and gestural expression but also aid in the coordination of each other's multimodal expression. In this way, people archive co-telling conversation in Japanese.

## References

1. Hayashi, M., Mori, J. and Takagi, T.: Contingent achievement of co-tellership in a Japanese conversation: An analysis of talk, gaze, and gesture. In C. Ford, B. Fox, and S. Thompson (eds.), *The Language of Turn and Sequence*, Oxford, Oxford University Press (2002) 81--122.
2. McNeill, D.: *Hand and Mind : What Gestures Reveal about Thought*, The University of Chicago Press, (1992)
3. Joh, A. and Hosoma, H.: Simultaneous Gestural Matching in Multi-Party Conversations . *Cognitive Studies*, Vol.16, (2009) 103--119 .
4. Sumi, Y., Yano, M., and Nishida, T.: Analysis environment of conversational structure *with nonverbal multimodal* data, The Twelfth International Conference on Multimodal Interfaces and the Seventh Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010), Beijing, China, November (2010)
5. Kendon, A.: *Gesture : Visible Action as Utterance*, Cambridge University Press, (2004)
6. Schegloff, E. A., Jefferson, G. and Sacks. H.: The preference for self-correction in the organization of repair in conversation, *Language*, Vol.53, No.2, (1977) 361—382
7. McNeill, D., Quek, F., McCullough, K., Duncan, S., Bryll, R., Ma, X. and Ansari, R.: Catchments, prosody and discourse, *Gesture*, Vol.1, No.1, (2001) 9—33
8. Heritage, J. & Raymond, G.: The terms of agreement: Indexing epistemic authority and subordination in assessment sequences. *Social Psychology Quarterly*, Vol.68, (2005) 15-38