Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme

Yasuharu Den* **Hanae Koiso†** **Takehiko Maruyama†**
Kikuo Maekawa† **Katsuya Takanashi‡** **Mika Enomoto§** **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,maruyama,kikuo,naou.yoshida}@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

§School of Media Science, Tokyo University of Technology
1404-1 Katakuramachi, Hachioji, Tokyo 192-0982, Japan
menomoto@media.teu.ac.jp

¶Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
4259 Nagatsutacho, Midori-ku, Yokohama 226-8503, Japan

Abstract

In this paper, we propose a scheme for annotating utterance-level units in Japanese dialogs, which emerged from an analysis of the interrelationship among four schemes, i) inter-pausal units, ii) intonation units, iii) clause units, and iv) pragmatic units. The associations among the labels of these four units were illustrated by multiple correspondence analysis and hierarchical cluster analysis. Based on these results, we prescribe utterance-unit identification rules, which identify two sorts of utterance-units with different granularities: *short and long utterance-units*. Short utterance-units are identified by acoustic and prosodic disjuncture, and they are considered to constitute units of speaker's planning and hearer's understanding. Long utterance-units, on the other hand, are recognized by syntactic and pragmatic disjuncture, and they are regarded as units of interaction. We explore some characteristics of these utterance-units, focusing particularly on unit duration and syntactic property, other participants' responses, and mismatch between the two-levels. We also discuss how our two-level utterance-units are useful in analyzing cognitive and communicative aspects of spoken dialogs.

1. Introduction

Unlike written text, spoken discourse does not exhibit sentence or utterance boundaries explicitly. Nonetheless, we have an intuition that there are some units that form segments at a certain level higher than words and phrases. Development of an established scheme for annotating such units is a crucial step towards corpus studies of spoken discourse and dialog.

Several attempts have been made to define utterance-units from various aspects including prosody (Du Bois et al., 1993; Beckman and Ayers, 1994), syntax (Meteer et al., 1995), and pragmatics (AMI, 2005). Yet, we do not have a widely-used scheme for identifying such units in dialogs.

Ford and Thompson (1996) analyzed the interrelationship among prosodic, syntactic, and pragmatic completion points of utterances in English conversations, showing that the majority of speaker changes occurred at *complex transition-relevance places*, which are defined by the convergence of prosodic, syntactic, and pragmatic completions. This result suggests that utterance-units suitable for dialog research should be defined in a complex way by taking all of prosody, syntax, and pragmatics into account.

In this paper, we propose a new scheme for annotating utterance-units in Japanese dialogs. We first apply three preexisting schemes, i) inter-pausal units, ii) intonation units, and iii) clause units, which have been adopted to annotation of dialog data, as well as one newly created scheme, iv) pragmatic units. We then analyze the interrelationship among the four units by using correspondence analysis and cluster analysis, showing that the labels of these units can be classified into several groups according to the depth of unit boundary. Based on these results, we come up with an annotation scheme that integrates the four schemes, distinguishing two sorts of utterance-units with different granularities: *short and long utterance-units*. Short utterance-units are identified by a pause and an intonation break, whereas long utterance-units are recognized by syntactic and pragmatic disjuncture. Applying this scheme to our dialog data, we explore some characteristics of our utterance-units, focusing particularly on unit duration and syntactic property, other participants' responses, and mismatch between the two-levels. We finally discuss how our two-level utterance-units are useful in analyzing cognitive and communicative aspects of spoken dialogs.

2. Analysis of the interrelationship among preexisting utterance-units

2.1. Data

Four dialogs from the *Chiba three-party conversation corpus* (Den and Enomoto, 2007) and another four di-

Table 1: Annotation labels

Inter-Pausal Unit (IPU)	
100	Followed by a pause longer than 100 msec
Intonation Unit (IU)	
3-H%	BI = 3, Tone = H%, LH%
3-HL%	BI = 3, Tone = HL%, LHL%
3-L%	BI = 3, Tone = L%
2p-H%	BI = 2 + pause, Tone = H%, LH%
2p-HL%	BI = 2 + pause, Tone = HL%, LHL%
2p-L%	BI = 2 + pause, Tone = L%
2-H%	BI = 2, Tone = H%, LH%
2-HL%	BI = 2, Tone = HL%, LHL%
2-L%	BI = 2, Tone = L%
F	BI = F
D	BI = D
Clause Unit (CU)	
AB	Absolute boundary
SB	Strong boundary
WB	Weak boundary constituting a CU boundary
NB	Non-predicative boundary
MB	Unit-initial/final interjection
FB	Unit-initial/final word fragment
Pragmatic Unit (PU)	
c	Communicative modality
e	Epistemic/deontic modality
n	Null modality
f	Unit-initial fragment
B	Backchanneling response token
E	Expressive response token
L	Lexical response token
O	Response token of other type (repetition, completion, or assessment)
Br	Reply/acknowledgment with B form
Er	Reply/acknowledgment with E form
Lr	Reply/acknowledgment with L form
Or	Reply/acknowledgment with O form

alogs from the *Corpus of Spontaneous Japanese (CSJ)* (Maekawa, 2003) were used for the current study. The Chiba dialogs were casual conversations among friends on campus, while the CSJ dialogs were dyadic conversations between interviewers and interviewees. For each dialog, a five-minute fragment, beginning one minute after the start of the dialog, was extracted for annotation and analysis. A total of 40-minute dialog fragments were used in the current study. All dialogs were carefully and precisely transcribed, including fillers, disfluencies, and laughter, and manually segmented into words with time information supplied at every word boundary.

2.2. Annotation

Three preexisting utterance-units, i) inter-pausal units (IPUs), ii) intonation units (IUs), and iii) clause units (CUs), as well as one newly created one, iv) pragmatic units (PUs), were identified in the data. The annotation of these units, except for IPUs, whose annotation was automatic, were performed by distinct three non-expert annotators, and crosschecked by at least one of the authors for each. Table 1 summarizes the annotation labels of these units.

Inter-pausal units (IPUs) IPUs (Koiso et al., 1998) were automatically identified by making reference to the time-stamps in the word-segmented transcriptions. A stretch of speech followed by a pause longer than 100 msec were recognized as an IPU.

Intonation units (IUs) IUs were labeled based on the X-JToBI scheme (Maekawa et al., 2002), an extension, for spontaneous speech, to the standard J_ToBI (Venditti, 1994). According to perceived intonational disjuncture, a break index (BI), indicated by a number, 1, 2, or 3, was assigned to every word boundary. When a boundary with BI = 2 was followed by a perceived pause, BI = 2p was used instead. A stretch of speech delimited by boundaries with BIs greater than or equal to 2 was recognized as an IU, which roughly corresponds to an accentual phrase. A final boundary tone, L%, H% (LH%), or HL% (LHL%), was also associated with each IU. Fillers, along with a certain set of interjections, and disfluencies were labeled with special marks, 'F' and 'D', respectively.

Clause units (CUs) Japanese is an SOV language, and one or more auxiliary verbs and conjunctive/final particles usually follow a predicate. Particularly, in colloquial Japanese, conjunctive particles are frequently used to iteratively concatenate a considerable number of clauses, which results in a very long 'sentence' without being accompanied by an explicit sentence final marker such as a conclusive form and a final particle (Iwasaki and Ono, 2002). Thus, some sort of morpho-syntactic criteria should be adopted to segment them into more tractable syntactic units.

CUs were originally designed to achieve such aim in segmenting monologs (Takanashi et al., 2003), and have been extended to cover dialog data. The scheme identified four types of CU boundaries: A(bsolute), S(trong), W(eak), and N(on-predicative) boundaries, which are characterized by explicit sentence final markers (AB), conjunctive particles expressing coordination (SB), other conjunctive particles followed by a discourse marker or speaker change (WB), and turn's completion with no predicate (NB), respectively. Two additional types for unit-initial/final interjections and word fragments were also used, i.e., M(iscellaneous) and F(ragmental) boundaries.

Pragmatic units (PUs) In addition to the three types of units described by acoustic, prosodic, and morpho-syntactic features, another kind of unit was also identified based on semantico-pragmatic properties. A PU was defined as a unit that constitutes a single proposition, except for that embedded in a relative clause or a propositional complement.

Linguistic modality, which refers to the speaker's mental attitude toward the propositional content and toward the hearer, was utilized to classify PUs. Three classes of linguistic modalities were distinguished: c(ommunicative), e(pistemic/deontic), and n(ull) modalities. Communicative modality included not only those expressed explicitly by grammatical devices such as *ne*, *yo*, and *ka* but also those expressed by rising intonation and implied by the context. Epistemic modality was expressed by predicate endings like *daroo*, *hazu-da*, and *no-da*, and deontic modality by expressions like *beki-da* and *nai-to-ike-nai*. When there were

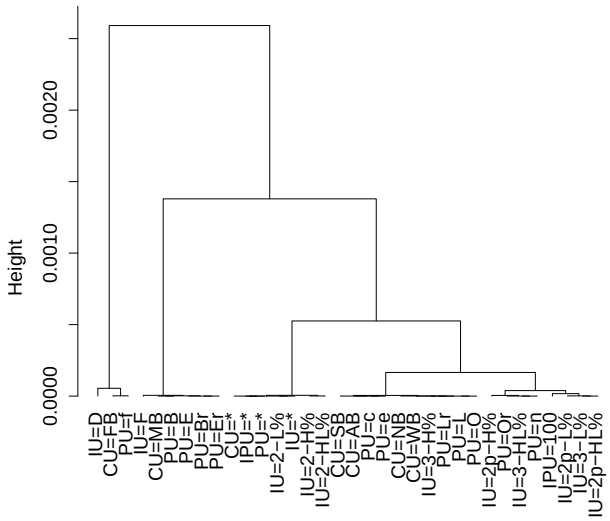

Figure 1: Cluster dendrogram for annotation labels

no such modality expressions, null modality was adopted.

An additional class was introduced to deal with f(ragments) due to false starts and self-interrupted speech. Furthermore, response tokens (Clancy et al., 1996; Gardner, 2001), which are produced by a hearer during a speaker's turn, were recognized separately, and classified into four types: B(ackchanneling) interjections such as *un* and *hai*, E(xpressive) interjections such as *a* and *hee*, L(exical) response tokens such as *soo* and *naruhodo*, and the O(ther) type, including repetitions of (part of) other's speech, collaborative completions, and assessments. When tokens with the same forms as these response tokens were used as replies to questions, requests, etc. or acknowledgments to them, they were labeled Br, Er, Lr, and Or.

2.3. Statistical analysis

The annotations of the four units above were combined into a tabular form, in which four types of annotation labels were aligned at every word. When a unit had no boundary at a given word, a special label '*', indicating 'no boundary,' was assigned. A total of 9266 words, 5001 in the Chiba dialogs and 4265 in the CSJ dialogs, were obtained and used in the subsequent statistical analysis.

In order to investigate the interrelationship among the four units, multiple correspondence analysis was, first, applied to the aligned annotation labels. Multiple correspondence analysis produces a geographic configuration of labels from multiple factors, in which labels with similar distributions are located close together. Hierarchical cluster analysis was, then, conducted to classify annotation labels based on the three-dimensional coordinates obtained by the multiple correspondence analysis. The distance measure was Euclidean, and the agglomeration method was Ward method.

2.4. Results

Figure 1 shows the cluster dendrogram for the annotation labels. It was evident that the labels could be classified into five groups. Major features of these groups are:

[Boundary classification rules]
Apply the following rules in this order at every word boundary:

1. If the tokens so far constitute a fragment, mark the boundary with 'F';
2. Else if the tokens so far constitute a backchanneling or expressive interjection, including one functions as a reply or an acknowledgment, mark the boundary with 'R';
3. Else if the current boundary exhibits a syntactic and/or pragmatic disjuncture, which may be expressed by a clause-unit boundary, a linguistic modality, or a turn-completing token, mark the boundary with 'L';
4. Else if the current boundary exhibits an acoustic and/or prosodic disjuncture, which may be expressed by a pause or an intonation break, mark the boundary with 'S';
5. Otherwise, apply these rules at the next word boundary.

[Unit identification rules]

Short utterance-units: Identify all four types of boundaries above as boundaries of short utterance-units.

Long utterance-units: With concatenating units labeled 'S' with the succeeding ones, identify the remaining boundaries as boundaries of long utterance-units.

Figure 2: Rules identifying short and long utterance-units

#1 Syntactic and pragmatic disjuncture
(CU=AB,SB,WB,NB; PU=c,e,L,O,Lr)

#2 Acoustic and prosodic disjuncture
(IPU=100; IU=3-,2p-)

#3 No boundary (IPU=*; IU=2-,*; CU=*; PU=*)

#4 Fragments (IU=D; CU=FB; PU=f)

#5 Backchanneling and expressive interjections
(IU=F; CU=MB; PU=B,E,Br,Er)

Three of these groups, #1, #2, and #3, appear to be in order of depth of boundary; syntactic and pragmatic disjuncture is deeper than acoustic and prosodic disjuncture, which is deeper than no boundary. They can be put on a spectrum according to the boundary-depth. The remaining groups, #4 and #5, seem better treated aside from the boundary-depth spectrum.

These results led us to an utterance-unit annotation scheme that integrates the four schemes we have discussed so far.

3. Proposal of two-level annotation scheme

Now, we are in a position to propose our empirically emerged two-level annotation scheme for utterance-units in Japanese dialogs. Syntactic and pragmatic disjuncture proposes deeper unit boundaries, whereas acoustic and prosodic disjuncture marks shallower boundaries. It would

Transcription			Gloss	IPU	IU	CU	PU	UU
120.08	120.71	R:kekkoo-ne	fairly-FP	100	3-H%	NB	c	L
120.85	121.16	L:nne	FP	100	3-H%	NB	L	L
121.16	121.42	R:un	yeah	*	F	MB	B	R
121.42	122.25	R:na//kanaka	rather	*	3-L%	*	*	S
122.29	123.89	R://hatu-kaigai-ryokoo-to-si-//te-wa	first overseas travel-as-TOP	100	3-L%	NB	n	L
121.66	121.88	L:(D in)	im-	100	D	FB	*	S
122.39	123.13	L:inpakuto-ga	impact-NOM	100	2p-H%	NB	n	L
123.57	124.77	L:<laugh>		-	-	-	-	-
124.86	125.52	L:naruhodo	I see	100	3-L%	NB	L	L
125.64	125.87	R:un	yeah	100	F	MB	B	R
126.15	126.28	L:de	and	*	3-L%	*	*	S
126.28	128.81	L:sono ano: tyuugoku-ni iku kikkake-n nat-ta-no-ga	uh uh China-DAT go opportunity-DAT become-PAST-N-NOM	*	3-L%	*	*	S
128.81	130.90	L:tyuutaa-o yat-te-ta-//tte-yuu-koto-//na-n-desu-kedo	tutor-ACC do-PAST-QP-thing-COP-N-POL-but	*	3-H%	SB	e	L
129.86	130.19	R:un	yeah	100	F	MB	B	R
130.43	130.70	R:un	yeah	100	F	*	*	S
130.86	131.09	R:hai	yes	100	3-L%	MB	B	R
130.90	131.09	L:sore	that	*	3-L%	*	*	S
131.09	131.68	L:daigaku-de	university-at	100	3-L%	NB	c	L
132.05	132.62	R:soo-desu	so-POL	100	3-L%	AB	Lr	L
132.88	133.31	L://(D n)<?>	n-	100	D	FB	f	F
132.90	133.14	R:un	yeah	100	F	MB	B	R
133.62	134.80	R:ano: daigaku-de	uh university-at	100	3-L%	*	n	S
135.00	136.25	R:ano daigaku-ttyuu-ka	uh university-QP-Q	100	3-L%	WB	n	L
136.53	136.92	R:(D s)soo	s- so	*	3-L%	NB	L	L
136.94	138.21	R:ano: tanom-are-te	uh ask-PASS-CP	100	3-L%	WB	n	L

Figure 3: Example of the annotation of the four units (IPUs, IUs, CUs, and PUs) as well as the proposed utterance-units (UUs). Each row corresponds to a short utterance-unit. Long utterance-units can be obtained by concatenating rows labeled 'S' with the succeeding rows. Numbers on the leftmost two columns indicate the starting and the ending times of the unit on that row, and the capital letter followed by '`:`' indicates the speaker. '`//`' means that a succeeding utterance by the other party begins overlapping at that location. The following glosses are used; NOM: nominative case marker, ACC: accusative case marker, DAT: dative case marker, TOP: topic marker, COP: copula, N: nominalizer, PASS: passive voice, PAST: past tense, POL: politeness marker, CP: conjunctive particle, FP: final particle, QP: quotative particle, Q: question marker.

be natural to assume that there is a hierarchical relationship between utterance-units determined by acoustics and prosody and those determined by syntax and pragmatics, the former being subsumed under the latter, although there may be a debate on this issue (see §4.3).

We, thus, define utterance-units bounded by acoustic and prosodic disjuncture as *short utterance-units (SUUs)*, and those bounded by syntactic and pragmatic disjuncture as *long utterance-units (LUUs)*. In addition, backchanneling and expressive interjections and fragments are identified separately, which are operationally included in both SUUs and LUUs. The procedures shown in Figure 2 enable us to recognize these utterance-units in dialogs.

Figure 3 depicts an example of our utterance-units, together with the underlying annotations of the four units.

4. Some characteristics of the proposed utterance-units

In this section, we explore some characteristics of our utterance-units, focusing particularly on unit duration and syntactic property, other participants' responses, and mismatch between short and long utterance-unit boundaries.

4.1. Unit duration and syntactic property

4.1.1. Purpose

The aim of this section is to examine the prosodic and syntactic properties of our utterance-units and to make clear what kind of units they are. We first analyze the distribution of the durations of SUUs, and show how they are related to units of speaker's planning. We then analyze the distribution of the word classes of the last words in SUUs, and discuss its implication with respect to turn-construction.

4.1.2. Data

For the dialog data described in §2.1, 3151 SUUs (Chiba: 1716; CSJ: 1435) and 1892 LUUs (Chiba: 1168; CSJ: 724) were identified by using the procedures shown in Figure 2. Of these data, only those LUUs labeled 'L', as well as the SUUs contained in them, were used in the current analysis.

4.1.3. Results and discussion

Table 2 shows the 0%, 25%, 50%, 75%, and 100% percentiles of the durations of the SUUs and LUUs in the Chiba and CSJ dialogs. The distributions for the SUUs were relatively narrow, with the inter-quantile ranges (IQRs) being 0.64 sec for the Chiba dialogs and 0.68 sec for

Table 2: Durations of short and long utterance-units (sec)

	N	0%	25%	50%	75%	100%
Chiba						
SUUs	1154	0.044	0.420	0.714	1.061	4.867
LUUs	617	0.166	0.632	1.107	2.089	15.280
CSJ						
SUUs	1063	0.036	0.445	0.734	1.126	3.101
LUUs	374	0.144	0.852	1.966	4.042	22.430

Table 3: Durations of short utterance-units (sec) relative to their locations in LUUs

	N	0%	25%	50%	75%	100%
Chiba						
Initial	240	0.050	0.251	0.473	0.804	3.329
Medial	297	0.061	0.364	0.635	1.005	3.647
Final	240	0.044	0.705	0.947	1.389	3.404
Single	377	0.166	0.486	0.730	1.068	4.867
CSJ						
Initial	210	0.054	0.326	0.572	0.952	3.101
Medial	479	0.036	0.400	0.677	1.058	2.607
Final	210	0.094	0.677	1.014	1.457	3.046
Single	164	0.144	0.540	0.811	1.098	2.327

the CSJ dialogs, compared with those for the LUUs, whose IQRs were 1.46 sec (Chiba) and 3.19 sec (CSJ). In addition, the medians for the SUUs in the two corpora were in accordance with each other, at about 0.7 sec. These findings suggest that SUUs may be results of some cognitive process inside the speaker that functions uniformly across speech situations, like Chafe's *idea units* (Chafe, 1994).

To look more closely at the duration of SUUs, the data for the SUUs shown in Table 2 were broken down into four sub-sets according to their locations in LUUs, as shown in Table 3. 'Initial,' 'Medial,' and 'Final' correspond to SUUs located at the initial, medial, and final locations in LUUs, respectively, and 'Single' corresponds to SUUs that solely constitute LUUs. Obviously, the durations of the final and single SUUs were longer than those of the initial and medial SUUs, suggesting that SUUs were not uniform within LUUs; the SUUs at LUU boundaries were construed as longer units than those at other places. Furthermore, the convergence of the distributions between the two corpora was evident in the medial SUUs, the IQRs being about 0.65 msec and the medians being about 0.65 msec.

The effect of the location was also observed in the syntactic property of SUUs. Table 4 shows the top three word classes of the last words in SUUs relative to their locations in LUUs. The final SUUs, and, hence, the LUUs containing them, often ended with final or conjunctive particles or auxiliary verbs (about 80% of the time), which is an expected feature of Japanese utterances (see §2.2). For the medial SUUs, on the other hand, case markers were the most frequent word class appearing at SUU boundaries, although their usage rate was not prominent. It is said that turn-construction in Japanese is advanced in an incremental fashion (Tanaka, 1999); case markers progressively project the turn-final shape, and utterance-final elements, such as auxiliary verbs and final particles, are placed after a clause-final predicate and thereby mark a possible completion point of the turn. In this respect, it may be stated that SUUs are building blocks for basic units of interaction, which are realized as LUUs, a similar perspective underlying the idea of *turn-constructional units (TCUs)* (Sacks et al., 1974; Schegloff, 1996).

Table 4: Top three word classes of the last words in SUUs relative to their locations in LUUs. CM: case marker, TM: topic marker, AP: adverbial particle, CP: conjunctive particle, FP: final particle, Aux.: auxiliary verb, CN: common noun, Adv.: adverb, Conj.: conjunction

	#1	#2	#3
Chiba			
Initial	Adv. (16.7%)	AP (11.7%)	CN (11.7%)
Medial	CM (15.5%)	Adv. (10.8%)	CP (9.8%)
Final	FP (42.5%)	CP (21.7%)	Aux. (13.3%)
Single	FP (32.6%)	Aux. (16.4%)	Adv. (10.6%)
CSJ			
Initial	Conj. (17.1%)	Adv. (14.8%)	CM (12.9%)
Medial	CM (17.5%)	Adv. (11.7%)	TM (10.6%)
Final	FP (40.0%)	CP (25.7%)	Aux. (14.3%)
Single	FP (46.3%)	Adv. (16.5%)	Aux. (15.9%)

4.2. Other participants' responses

4.2.1. Purpose

The aim of this section is to examine how an utterance-unit being produced by a speaker is treated by other participants, by analyzing the timing of other participants' responses to SUUs and LUUs. We suppose that LUUs constitute basic units of interaction, and, thus, predict that speaker transition would be localized at LUU boundaries. We also suppose that SUUs are not only units of speaker's planning but also units of hearer's understanding. Thus, we predict that boundaries of SUUs would provide opportunities for backchanneling and expressive response tokens, which are considered as signals of hearer's understanding and change of hearer's mental state.

4.2.2. Data and annotation

In order to distinguish some distinct patterns of speaker transitions, according to the ways of the progress of conversation and to turn-taking rules, the following turn-transition tags were assigned to the LUU data used in §4.1.

First, each dialog was segmented into several chunks, each of which was classified into either a turn-by-turn or telling stage; in the former stage, utterances are produced in turn by two or more speakers, following the turn-taking system (Sacks et al., 1974), whereas in the latter stage, a single speaker, telling a story or giving an explanation, exclusively keep a turn, others supporting his/her multi-unit turn as recipients. Then, for each LUU at a turn-by-turn stage, its antecedent unit was identified by making reference to the time information and the content of the utterance, and the current unit was classified into three types according to the

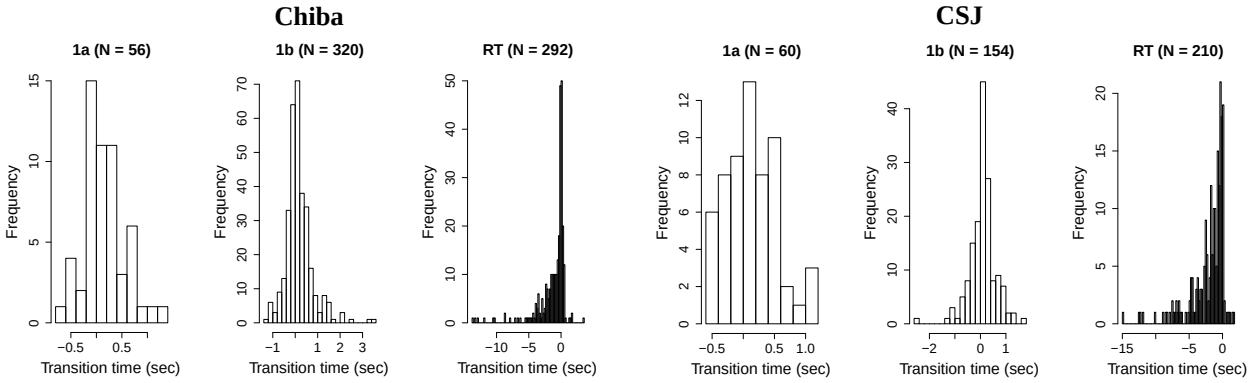

Figure 4: Distributions of transition times between adjacent long utterance-units relative to turn-transition types. The histograms for the speaker-continuation types (1c and s) were omitted. The histograms entitled 'RT' show the cases where second units are backchanneling or expressive response tokens. In each histogram, the width of a cell is fixed to 200 msec.

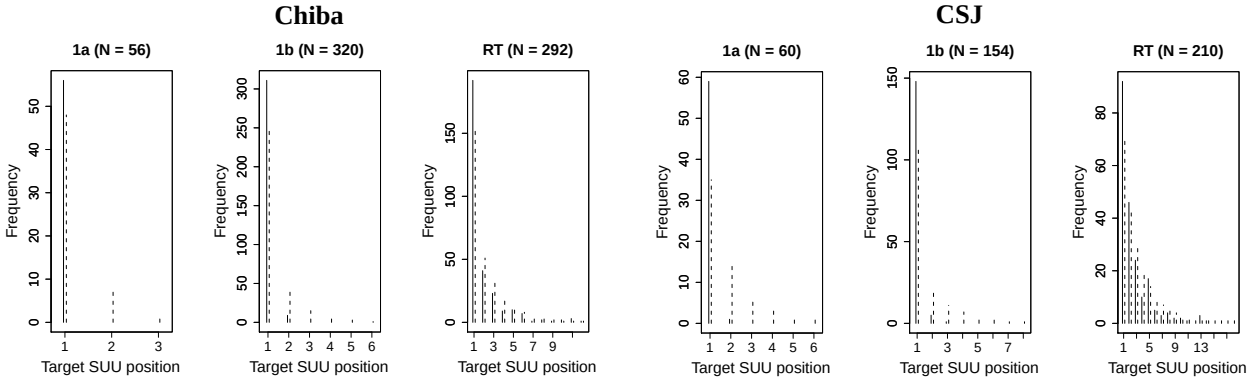

Figure 5: Distributions of target SUU positions measured from the end of the antecedent unit relative to turn-transition types. Solid lines represent the observed distributions, whereas broken lines represent the distributions predicted by a random model in which the target SUU was selected from each antecedent unit with equal probability.

turn-taking rules being employed (Sacks et al., 1974).

1a The current speaker has been selected as next speaker by means of a next-speaker-selection technique, utilized by the speaker of the antecedent unit, such as the affiliation of an address term or a gaze at one party to a class of utterances such as question, request, etc.

1b The current speaker has selected himself/herself as next speaker, being the first to start a new turn.

1c The speaker of the antecedent unit has continued his/her turn.

Continuation of a telling sequence by the primary speaker at a telling stage was separately labeled as 's'. For response tokens and fragments occurring within other participants' utterance-units, no turn-transition tag was assigned.

The annotation was conducted by one of the authors. In annotating with these tags, the annotator ignored whether or not the current unit was properly launched at the transition-relevance place (TRP) of the antecedent unit. This is because, if we considered only those utterance units that started at TRPs, the timing of turn-taking would be artificially localized at the end of utterance-units, leading us to *petitio principii*.

4.2.3. Results and discussion

Figure 4 shows the histograms of transition times between adjacent LUUs relative to turn-transition types. The cases where second units were backchanneling or expressive response tokens (RTs) are also included.

The ratios of speaker-change types (1a and 1b) to speaker-continuation types (1c and s) in the Chiba and CSJ dialogs were 45.7% (376:447) and 45.9% (214:252), respectively. This means that about a half of the LUUs were accompanied by other participants' start of a new turn.

In the 1a and 1b data, we found that the peaks of the distributions were all located at $-200 \sim 0$ msec or $0 \sim 200$ msec and that 95% of the data fell within the range of about 1.5 sec in the 1a data (Chiba: 1.4 sec; CSJ: 1.6 sec) and the range of about 2.3 sec in the 1b data (Chiba: 2.4 sec; CSJ: 2.2 sec). These values contrasted with that in the RT data, which was about 9 sec (Chiba: 9.0 sec; CSJ: 9.1 sec).

In order to see the relation of other participants' responses to SUUs, the target SUU in the antecedent unit was also identified for each adjacent LUU pair based on the time information; that is, the target SUU was defined as the last SUU whose ending time was not beyond the starting time of the current unit. The solid lines in Figure 5 show the observed distributions of the positions of target SUUs mea-

sured from the end of the antecedent unit for the 1a, 1b, and RT data. The broken lines, on the other hand, show the distributions predicted by a model in which the target SUU was selected from each antecedent unit with equal probability. As clearly seen in the 1a and 1b data, virtually all responses occurred at the final SUU in the antecedent unit, although, in theory, earlier SUUs also could have been the target. A dramatic difference, however, was observed in the RT data, where the observed and the predicted distributions were rather similar.

In sum, the timing of turn-taking was localized at LUU boundaries, suggesting the adequacy of LUUs as units of interaction. In contrast, the chance of eliciting a response token was nearly equal for all SUUs contained in an LUU, suggesting that SUUs may be well suited for units of hearer's understanding.

4.3. Mismatch between short and long utterance-unit boundaries

4.3.1. Purpose

In prescribing the utterance-unit identification rules shown in Figure 2, we have assumed hierarchical relationship between SUUs and LUUs; in other words, we have presupposed that boundaries labeled 'L' constitute not only LUU boundaries but also SUU boundaries. There were, however, a few cases where LUU boundaries identified by rule 3 did not possess the properties of SUUs characterized by rule 4. The aim of this section is to focus on these LUU boundaries, which involve *mismatches* with SUU boundaries, and to consider their meaning.

4.3.2. Data

From among the 991 LUUs used in §4.1, 64 instances (Chiba: 52 (= 8%); CSJ: 12 (= 3%)) were extracted, which did not have the property of SUU boundaries, i.e., `IPU=100` or `BI=3-,2p-`. These instances were classified into several patterns according to the syntactic and pragmatic contexts in which they occurred. For some of them, another set of instances that appeared in similar contexts but that exhibited no mismatch were also extracted and compared with the mismatch cases.

4.3.3. Results and discussion

The majority of mismatch instances could be classified into the following patterns. In the following examples, each line corresponds to an LUU and the mismatch boundary is marked with labels '`[IPU,IU,CU,PU]`'.

1. Inversions (18 cases)
   The mismatch boundary is immediately followed by an inverted (postposed) element.

```
A:Ii-too-no   sit-teru: [*,2-H%,*,c]
   E-building-GEN  know
   Do you know the one in building E?
A:ano koohii-meekaa
   that  coffee-machine
   That coffee machine.
```

2. Prefaces (6 cases)
   The LUU in question is a preface to the body of the speaker's turn, projecting the continuation of his/her turn across the mismatch boundary.

```
A:itumo omou-n-dakedo [*,2-L%,SB,e]
   always  think-N-but
   I always think about this, but
A:X-san-tte Y-san-ni-taisi-te
   X-Ms.-QP    Y-Mr.-to
   tyotto-sa: kekkoo: yuu-yo-ne
   just-FP         much       say-FP-FP
   Ms. X just say much to Mr. Y, doesn't she?
```

3. Lexical response tokens (10 cases)
   The LUU in question is a lexical response token *soo* or *soo-desu(-ne)*, which is immediately followed by a substantial utterance by the same speaker, yielding a resumptive opener (Clancy et al., 1996).

```
C:sakazuki-mitai-na-no
   cup-like-COP-FP
   Is it like a cup?
A:soo [*,2-L%,NB,Lr]
   yes
   Yes,
A:de sore-no repurika-to-ka
   and  it-GEN     replica-or something
   ut-teru-tte
   sell-QP
   and they sell its replica or something.
```

4. Repeats of predicates (4 cases)
   The LUU in question is repeated immediately afterward for the purpose of emphasis, etc.

```
A:it-teru [*,2-L%,AB,n]
   hiss
   It's hissing,
A:it-teru
   hiss
   hissing.
```

At these mismatch boundaries, other participants rarely started their new turn; the rate was 27% in the Chiba dialogs and no start of other's turn was observed in the CSJ dialogs. One may conjecture that in these cases other participants' responses were suppressed by the current speaker's use of some acoustic and/or prosodic techniques that enabled hearers to predict speaker continuation. However, it is not necessarily the case. In these syntactic and pragmatic contexts, the response rate was lower than in other contexts, even when there was a pause or an intonation break, i.e., no mismatch was concerned. (For instance, only about 30% of inversions and about 12% of lexical response tokens involved other participants' responses, which were less than the overall response rate of about 46%.)

It would be more plausible that the syntactic and pragmatic factors involved in these contexts provide hearers with cues for speaker continuation regardless of whether or not they are accompanied by acoustic and prosodic devices. For instance, a lexical response token *soo* is used not only to display understanding but also to acknowledge that the previous utterance by other party has represented the speaker's own opinion and, thus, *soo* can serve as a preface to the body of his/her turn where his/her opinion is to be developed based on the other party's previous contribution (Kushida, 2002). In such circumstances, hearers may naturally expect the continuation of the current turn after *soo*.
At this time, we have not come up with specific meaning of these mismatches. However, they provide some motivation for us to consider more carefully and deeply the implications of our utterance-units. In addition, the presence of mismatches may lead us to reconsider our prerequisite of hierarchical relationship between SUUs and LUUs.

5. General discussion

Finally, we discuss implications of our annotation scheme and future plans.
SUUs are identified by acoustic and prosodic features. A pause and an intonation break are decisive cues to recognize these units. This is consistent with Chafe's notion of *idea units*, which have been considered to be one of the most important concepts in spoken discourse studies from a cognitive view point (Chafe, 1994). Some characteristics of SUUs shown in §4 also supported this correspondence. Chafe (1994) suggests that each idea unit represents the information that is active in the speaker's mind at the moment in discourse. In this respect, SUUs would be useful in the study of cognitive aspect of spoken dialogs.
LUUs, on the other hand, are determined by syntactic and pragmatic features. Syntactic and pragmatic completions play an important role in recognizing LUUs. Since most LUU boundaries exhibit prosodic completion, LUUs are seen as units defined by the convergence of prosodic, syntactic, and pragmatic completions. This is parallel to *complex turn-constructional units*, which have been proposed as basic units of interaction in the conversation analysis literature (Ford and Thompson, 1996). Some characteristics of LUUs shown in §4 suggested the adequacy of our LUUs as units of interaction. Thus, LUUs would be useful in the study of communicative aspect of spoken dialogs.
To promote these lines of research, we are planning to develop an annotation scheme for *functions* of our utterance-units. For the study of cognitive aspect, information status of discourse elements, such as 'given' and 'new,' may be of use. We are investigating direction to such functional annotation of SUUs. For the study of communicative aspect, on the other hand, it is fundamental to represent structures of turns and actions implemented therein. We are trying to decide on a scheme, which is applicable not only to dyadic conversations but also to multi-party conversations, to represent such structures out of LUUs.

6. References

The AMI Project, 2005. *Guidelines for dialogue act and addressee annotation version 1.0*.

M. Beckman and G. Ayers, 1994. *Guidelines for ToBI labeling*. Ohio State University.

W. L. Chafe. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press, Chicago.

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.

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.

J. W. Du Bois, S. Shuetze-Coburn, S. Cumming, and D. Paolino. 1993. Outline of discourse transcription. In J. A. Edwards and M. D. Lampert, editors, *Talking data: Transcription and coding in discourse research*, pages 45–89. Lawrence Erlbaum Associates, NJ.

C. E. Ford and S. A. Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff, and S. A. Thompson, editors, *Interaction and grammar*, pages 134–184. Cambridge University Press, Cambridge.

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

S. Iwasaki and T. Ono. 2002. "Sentence" in spontaneous spoken Japanese discourse. In J. Bybee and M. Noonan, editors, *Complex sentences in grammar and discourse: Essays in honor of Sandra A. Thompson*, pages 175–202. John Benjamins, Amsterdam.

H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. 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.

S. Kushida. 2002. Kaiwa no naka no "un" to "soo": Wasyasei to no kakawari de *("Un" and "soo" in conversation)* (in Japanese). In T. Sadanobu, editor, *"Un" to "soo" no gengogaku*, pages 5–46. Hitsuji-Syobo, Tokyo.

K. Maekawa, H. Kikuchi, Y. Igarashi, and J. J. Venditti. 2002. X-JToBI: An extended J_ToBI for spontaneous speech. In *Proceedings of the 7th International Conference on Spoken Language Processing (INTERSPEECH 2002)*, pages 1545–1548, Denver, CO.

K. Maekawa. 2003. Corpus of spontaneous Japanese: Its design and evaluation. In *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pages 7–12, Tokyo.

M. Meteer et al. (revised by A. Taylor), 1995. *Dysfluency annotation stylebook for the Switchboard corpus*.

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

E. A. Schegloff. 1996. Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, and S. A. Thompson, editors, *Interaction and grammar*, pages 52–133. Cambridge University Press, Cambridge.

K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara. 2003. Identification of "sentences" in spontaneous Japanese — detection and modification of clause boundaries —. In *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pages 183–186, Tokyo.

H. Tanaka. 1999. *Turn-taking in Japanese conversation: A study in grammar and interaction*. John Benjamins, Amsterdam.

J. J. Venditti. 1994. Japanese ToBI labelling guidelines. Technical Report 50, Ohio State University.