Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 51-54.

An Annotation Scheme for Syntactic Unit in Japanese Dialog

Takehiko Maruyama[1], *Katsuya Takanashi*[2], *Nao Yoshida*[1*]

[1]National Institute for Japanese Language and Linguistics, Japan
[2]Academic Center for Computing and Media Studies, Kyoto University, Japan
maruyama@ninjal.ac.jp, takanasi@ar.media.kyoto-u.ac.jp, naou.yoshida@ninjal.ac.jp

Abstract

In this paper, we propose a scheme for annotating syntactic units called **DCU** (Dialog Clause-Unit) in Japanese dialogs. Since there is no explicit devices to mark sentence boundaries in speech, precise definition and criteria must be designed to extract syntactic units from the utterance. We show a design of DCU which consists of clausal and non-clausal units. Annotating DCU tags to eight dialogs of 40 minutes from two different dialog corpora, we examine characteristics of each dialog from the viewpoint of DCU, and compare them to the distribution of clausal-units annotated to monologs.
Index Terms: *Dialog Clause-Unit*, Japanese dialog and monolog, clause boundary, unit length

1. Introduction

It has been broadly recognized that the most standard syntactic unit for various linguistic analyses, especially for syntax, is a "sentece." In written text, sentence boundaries are generally marked by punctuations, and thus it is easy to specify an extent of each sentence. In spoken language, however, there is no explicit devices to mark sentence boundaries, and so precise criteria must be defined in order to segment the flow of utterance into syntactic units. Moreover, dialogs frequently contain utterances with single-word, syntactic incompleteness, interjections, repetitions, repairs, inversions, insertions and so on, which often cause difficulties to define a certain syntactic entity. Therefore, definite procedure must be designed in each language to extract general syntactic units of dialog.

In this paper, we propose an annotation scheme for Japanese syntactic unit of dialogs, called "**DCU (Dialog Clause-Unit)**." We introduce a scheme for annotating clausal and non-clausal boundaries to segment the utterance into DCU. Since Japanese is an SOV language, predicates (verb phrases, for example) are placed at the end of each clause, and we can specify the type of each clause by their conjugated forms and conjunctive particles at the end of verb phrase. Using such morpho-syntactic clues, we define a procedure to annotate Japanese clause boundaries where utterance should be segmented into clausal units. Also we illustrate some linguistic phenomena which cause problems to specify appropriate syntactic units, which result in non-clausal units.

First we survey some preceding works that define basic unit for spontaneous speech from various aspects, including "**CU (Clause-Unit)**" which was originally designed for spontaneous monologs in *Corpus of Spontaneous Japanese* [1]. Second we describe the system of Japanese clause boundaries and their annotation scheme. Then we show the result of our current study, an annotation to a total of eight dialogs for 40 minutes from two different dialog corpora. We examine characteristics of each dialog from the viewpoint of DCU, and also compare them to the result of CU annotated to monologs.

[*]Until March, 2010.

2. Background

There have been various proposals to define basic unit of dialog from different aspects.

Chafe (1987) proposed **IU** (Intonation Unit) [2] which split utterances referring their intonation, and it has been broadly used in phonetic, phonological, and discourse analyses. Koiso et.al (1998) introduced **IPU** (Inter-Pausal Unit) [3], which a stretch of speech followed by a pause longer than 100 msec is recognized as a unit. Neither IUs nor IPUs, however, consists of syntactic units necessarily. **TCU** (Turn Constructional Unit) by Sacks (1974) split utterances in dialogs where turn-taking can occur [4]. Syntactic feature of the utterance may influence an extent of each turn, the criteria of TCU do not necessarily focused on syntactic viewpoint. Moreover, the annotator must consider the mechanism of turn-taking in the conversation. **AS-unit** (Analysis of Speech Unit) by Foster et.al (2000) is mainly a syntactic unit [5]. So far as clause boundaries works as crucial cues to segment the utterance and thus each unit consists a syntactic entity, aims of AS-unit and DCU are in common. In case of *Corpus of Spontaneous Japanese*, **CU** (Clause-Unit) has been proposed by Takanashi et.al (2003), which was designed and annotated as a syntactic unit of spontaneous monolog [6]. Each CU was used as a basic unit for discourse-level annotations; dependency structures, sentence extraction, and discourse structure annotation.

CU was originally designed for monolog, and we extend the criteria of CU to extract a basic syntactic unit for dialog, called **DCU** (Dialog Clause-Unit).

3. Annotation scheme for DCU

3.1. Clausal and Non-clausal Unit in Monolog and Dialog

Japanese morphological structure of predicates (verb phrases, adjective phrases, copulas and so on) has highly been developed by its conjugation form and/or conjunctive particle. Referring to these morpho-syntactic information, we can specify a boundary of clause and its syntactic feature quite precisely. In monologs narrations basically consists of clause linkages, including explicit sentence-final boundaries, as long as the speaker tries to speak tidily. In that respect, the utterance in monolog consists of a chain of clausal units.

In dialog, on the other hand, not only clausal units but also non-clausal entities are frequently observed, like one-word utterances, interjections, elliptical segments of phrases and so on. Dealing with this characteristic of dialog, we should positively admit and annotate non-clausal boundaries in DCU.

time stamp		spk.	utterance
337.8530	339.2690	A:	*ja hakkou si tari nai mama tabe te n da* /AB
338.9240	339.0340	B:	(D_*u*) /FB
339.2630	340.5100	C:	(I_*un*) (I_*a*) /MB *sore da:* /AB
339.5430	341.9600	B:	*uti tabe owat te kara gyuunyuu tasi te hitoban oi te ru kara sa:* /WB
342.1190	342.8750	C:	(I_*un*) (I_*un*) (I_*un*)
342.9000	344.5950	C:	(I_*a*) /MB *son kurai yan* (W_*nake*|*nakya*) (D_*mu*) *muri na no ka na* /AB
343.5120	344.4270	A:	*sou da yo ne:* /AB
343.9710	346.1840	B:	(I_*un*) /MB {Q *sore gurai no hindo no hou ga ii* |AB}+ *to omou yo* /AB

Figure 1: Example of an annotating DCU tags

3.2. Basic Scheme of Annotation

Tags of DCU are annotated on dialog transcriptions. An annotator reads the transcription listening to its original speech, and annotates **Clause Boundary Labels** at the end of clausal and non-clausal boundaries. Clausal and non-clausal boundaries to be annotated are previously defined, as shown in section 3.3.

After annotating clause boundary labels, manual adjustment is required for the phenomena of quotation, inversion, insertion, and try-marker in the middle of utterance. Then the annotator modifies the clause boundary labels appropriately by the prescribed procedure shown in section 3.4 to adjust them as suitable syntactic units.

Figure 1 shows an example of annotation. For example, "/AB" and "{Q *** |AB}+" correspond to clause boundary label and the result of manual modification, respectively.

3.3. Annotating Clause Boundary Labels

Clause boundary labels in DCU can be classified into six categories according to their syntactic features.

3.3.1. Absolute Boundary (/AB *)*

Absolute boundaries, annotated by a label "/AB ", should be annotated at the end of explicit boundaries of declarative, interrogative, imperative, and exclamatory statements (in other word, the end of "sentences"). Even if the utterance consists of only one word like *'ita'* ("there he was") or *'sugoi'* ("great!"), /AB should be annotated if it works as a predicate and functions as declarative or exclamatory statement.

3.3.2. Strong Boundary (/SB *)*

Strong boundaries, by a label "/SB ", should be annotated at the end of coordinate clauses, marked by conjunctive particles *ga*, *keredomo*, *keredo*, *kedomo*, *kedo* and *si*. These are clauses that are highly independent of their main clauses [7], and tend to form an independent statement, and thus should be treated as isolated syntactic units.

3.3.3. Weak Boundary (/WB *)*

Weak boundaries, by a label "/WB ", should be annotated at the end of subordinate clauses, marked by conjunctive particles *te*, *kara*, *noni*, *node*, *mitaina* and so on, only when the same speaker continue his/her speech begun with a conjunctive, or turn taking occurs after them. A conjunctive following a subordinate clause tends to cut off a chain of utterance clearly, and turn-taking after a subordinate clause means his/her utterance is forced to end (or just ends) at the clause boundary.

3.3.4. Non-predicative Boundary (/NB *)*

Whereas these three boundaries shown above are all clausal units, non-clausal units must also be annotated in dialog. Non-predicative boundaries, marked by "/NB ", corresponds to major breaks of utterance without predicates. It includes utterances with one or a few words like *'juunen'* ("ten years"), elliptical segments of phrases like *'daigaku dokono'* ("university, where?"), vocatives and interruptions. Boundaries after lexical interjections like *'honto'* ("really?") are also classified here.

3.3.5. Interjection Boundary (/IB *)*

If an interjection (transcribed by (I_***) in Figure 1) appears after a clausal or non-clausal boundary, it should be marked by "/IB " to be cut off from the rest of units. A succession of two or more interjections like '(I_*ee*)(I_*ee*)' ("yeah, yeah") should be treated together, as '(I_*ee*)(I_*ee*) /IB '.

3.3.6. Fragmental Boundary (/FB *)*

If a word fragment (transcribed by (D_***) in Figure 1) appears after clausal or non-clausal boundary, it should be marked by "/FB " to be cut off from the rest of units. A succession of two or more fragments like '(D_*u*)(D_*tyo*)' should be treated together, like '(D_*u*)(D_*tyo*) /FB '.

3.4. Modifying Clause Boundary Labels

After annotating the clause boundary labels, manual investigation and modification is required, because sometimes the label is not appropriate to be a boundary of a syntactic unit. The modification is applied in the case of quotation, inversion, insertion, and try-marker in the middle of utterance.

3.4.1. Quotation

An absolute boundary, for example, is sometimes not a suitable boundary of DCU, when it is quoted as shown in (1).

(1) *sugoi na:* /AB *tte omou* /AB
"it's great" "I think that"

Since *'tte'* is a quotation marker which quotes the preceding part with /AB boundary, two units in (1) should be united as a whole. If a clause boundary label is embedded in a quotation, the extent of quotation must be bracketed by {Q *** }+, and the label must be modified like "|AB" as follows, where "|" indicates that it is not a boundary of DCU.

(1') {Q *sugoi na:* |AB}+ *tte omou* /AB
"I think that it's great"

3.4.2. Inversion

In spontaneous speech an argument is sometimes irregularly placed after the predicate to which it is dependent syntactically. When such an inversion occurs over a clause boundary, the label must be modified like "|AB+" and inverted element(s) bracketed by "≪" and "≫." In this case, there must be a boundary after the inverted element(s) as /NB .

(2) *hidee na* /AB *sore* /NB
"terrible" "it's"

(2') *hidee na* |AB+ ≪*sore*≫ /NB
"it's terrible"

3.4.3. Insertion

A speaker sometimes inserts an independent syntactic chunk in the middle of the utterance, typically when he/she hesitates. The syntactic boundary at the end of hesitating part must be modified as non-boundary of DCU.

(3) *ato wa chotto* *nan daro:*/AB *syosinmono dakara*
"and a little bit" "let me see" "I'm coward"

As the speaker hesitates, *'nan daro:'* is inserted in the middle of the ongoing syntactic unit. In this case the label /AB must be modified as "|AB" and the inserted extent must be bracketed as {I ***}+.

(3') *ato wa chotto* {I *nan daro:*|AB}+ *syosinmono dakara*
"and, let me see, I'm coward a little bit"

3.4.4. Try-marker

When a speaker is not confident what he/she is about to say, the element is sometimes focused by interrogative form with /AB to mark the uncertainty, and the speaker keeps on uttering his/her speech without interruption after the boundary. In this case (see [8] "try-marker"), the annotated label /AB must be modified as "|AB+T."

(4) *juunenme gurai kana* /AB naru hito ga inno /AB
"about ten years, isn't it?" "there is a person"

(4') *juunenme gurai kana* |AB+T *naru hito ga inno* /AB
"there is a person (who works) for about ten years"

4. Analysis

4.1. Data

In this section we show the result of our current study annotating DCU to eight dialogs extracted from two different dialog corpora.

CSJ: *Corpus of Spontaneous Japanese* (Dialog part) [1]
– interviews in formal style.
– extracted four dialogs, a total of 4,302 words.

Chiba: *Chiba three-party conversation corpus* [10]
– casual conversations among three friends on campus.
– extracted four dialogs, a total of 5,201 words.

All dialogs have been carefully transcribed including tagged fillers, word fragments and interjections, with begin/end time and speaker information. Also they have been manually segmented into words. We used a five-minute fragment of dialog, beginning one minute after the start. Four fragments of dialog from each corpus, a total of 40 minutes, 9,503 words were extracted for annotation. Two annotators conducted DCU labeling along the scheme shown in section 3.

4.2. Distribution of Clause Boundaries Labels

A total of 734 clause boundary labels were annotated to CSJ, and 1,151 labels to Chiba. Figure 2 shows the distribution of clause boundary labels.

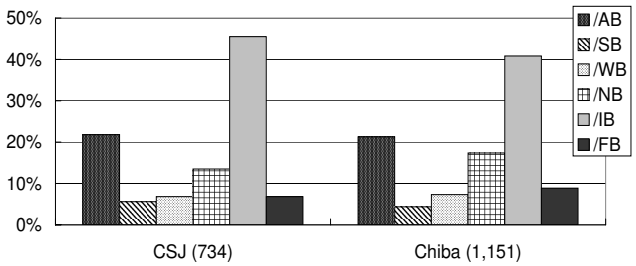

Figure 2: Distribution of Clause Boundary Labels

Both in CSJ and Chiba, /IB is the most frequent among the labels. This result shows that the proportion of DCU with isolated interjection which works as a backchannel is very high, which is typical and remarkable characteristic of dialogs compared to monologs, since interjections are not essentially included in the latter. Comparing the other labels, there seems no conspicuous difference between CSJ and Chiba.

4.3. Frequencies of Modification

Figure 3 shows the frequencies of modification.

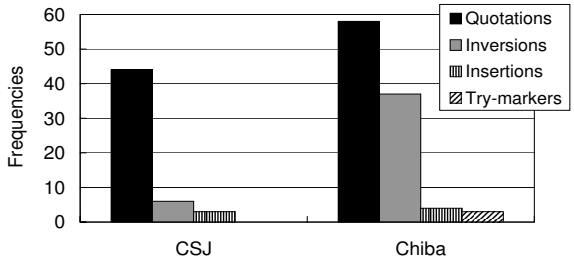

Figure 3: Modification of Clause Boundary Labels

The frequency of inversion in Chiba is much higher than that in CSJ. Since the participants in Chiba talk whatever comes to mind on a given occasion, the formality of speaking style is much lower, and spontaneity much higher than those of CSJ, which consists of formal interviews. Such a difference between the two dialogs might influence the frequencies of inversions.

4.4. Length of DCU

Segmenting the utterance at the end of annotated (and modified) clause boundary labels, we get the fixed extent of DCU. Tables 1 and 2 show the length of DCU, measured by the number of words within each unit.

The average length of DCU in CSJ tends to be longer than that in Chiba. In CSJ narratives by interviewees often appear, while in Chiba participants exchange short messages in casual conversations, and such difference make the average length of DCU longer in CSJ.

As shown in the tables, the mode length of /AB is three words in CSJ and four words in Chiba. Examining the data we found most of these examples were *'sou desu ne* /AB *'* ("that's right") or *'sou na n da* /AB *'* ("really"), a kind of fixed forms of agreement or confirmation. Furthermore, short units with /AB like *'tigau* /AB *'* ("it is not so") or *'wakan nai* /AB *'* ("I don't know") often appeared. These short units tend to function as

Table 1: Length of DCU in CSJ

	Average	Median	IQR	Max	Mode
/AB	11.3	9	11	77	3
/SB	19.8	14	14	63	14
/WB	13.0	11.5	7	35	11
/NB	6.1	3	4.5	39	2
/IB	1.3	1	0	7	1
/FB	1.1	1	0	2	1

Table 2: Length of DCU in Chiba

	Average	Median	IQR	Max	Mode
/AB	7.5	5	5	47	4
/SB	11.3	9.5	5.75	31	9
/WB	12.6	10	9	51	7
/NB	4.8	2	5	57	1
/IB	1.4	1	0	7	1
/FB	1.3	1	0	5	1

second pair parts of adjacency pairs or be placed as sequence-closing thirds [9], which is characteristic for dialog.

4.5. Comparison of Dialogs and Monologs

Finally, annotated DCU in dialogs are compared to CU in monologs. As noticed above, original CU was annotated to CSJ monolog part, a total of 177 monologs. Figure 4 shows the distribution of clause boundary labels in CSJ (dialog part), Chiba, and CSJ (monolog part), except /IB and /FB .

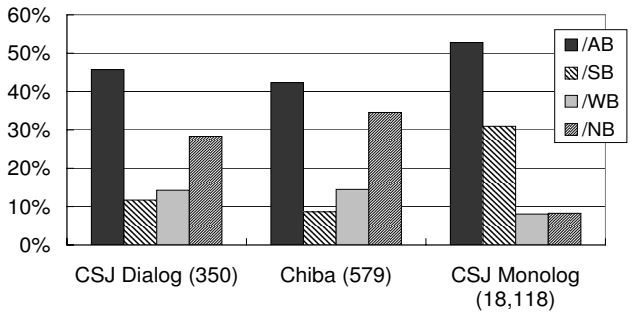

Figure 4: Clause Boundary Labels in Dialogs and Monolog

It is obvious that the total ratio of /AB and /SB is much higher in monologs than those in dialogs, while /WB and /NB much higher in dialogs than monologs. This contrast may be caused by the difference of speaking strategies between the two; in monologs speakers basically keep his/her narration with a sequence of explicit sentences or coordinate clauses, while in dialogs elliptical segments of phrases or non-clausal units tend to constitute short turns frequently.

Table 3 shows the length of CU, measured by the number of words within each unit.

Table 3: Length of CU in CSJ (monolog part)

	Average	Median	IQR	Max	Mode
/AB	28.1	24	23	147	17
/SB	21.5	18	15	148	13
/WB	23.5	19	21	115	9
/NB	13.5	10	12	111	2

Comparing the Tables 1, 2 and 3, the average and mode length of /AB in monolog are much longer than those in dialogs. This indicates that almost all the utterances in monolog consist of narratives which makes each unit quite long, while a lot of short unit with /AB (see section 4.4) is involved in dialogs.

As for the average length of /SB , on the other hand, there seems to be no difference between CSJ monolog and dialog. This is because interviewees in dialogs tend to reply to interviewer's questions using narrative forms which include a sequence of coordinate clauses marked by /SB .

5. Concluding Remarks

It has been broadly recognized that some general units should be designed appropriately according to each purpose of analyzing spontaneous speech [11]. In this paper we proposed an annotating scheme for DCU, a syntactic unit of dialog in Japanese. As shown in many preceding researches, clausal and non-clausal units are valid and effective from the viewpoint of syntax as well as information structure. Such units should be designed in each language considering its grammatical feature, which makes it possible to compare syntactic features between different languages directly. Moreover, it is expected to examine a relationship between syntactic and prosodic units, and a relationship between spoken and written language from the perspective of syntactic unit.

6. References

[1] Maekawa, K., "Corpus of spontaneous Japanese: Its design and evaluation", Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 7–12, 2003.

[2] Chafe, W., "Cognitive constraints on information flow", In R. Tomlin [Ed], Coherence and grounding in discourse, John Benjamins, 1987.

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

[4] Sacks, H., Schegloff, E. A. and Jefferson, G., "A simplest systematics for organization of turn-taking for conversation", Language, 50(4):696–735, 1974.

[5] Foster, P., Tonkyn, A. and Wigglesworth, G., "Measuring spoken language: A unit for all reasons", Applied Linguistics, 21:354–375, 2000.

[6] Takanashi, K., Maruyama, T., Uchimoto, K. and Isahara, H., "Identification of "Sentence" in Spontaneous Japanese – Detection and modification of clause boundaries –", Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 183–186, 2003.

[7] Minami, F., *Gendai nihongo no kouzou* (Structure of Modern Japanese), Taishukan shoten, 1974.

[8] Sacks, H. and Schegloff, E.A., "Two Preferences in the Organization of Reference to Persons in Conversation and Their Interaction", in G. Psathas [Ed], Everyday Language: Studies in Ethnomethodology, 15–21, Irvington Press, 1979.

[9] Schegloff, E. A., Sequence Organization in Interaction: A primer in Conversation Analysis 1, Cambridge University Press, 2007.

[10] Den, Y. and Enomoto, M., "A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation", in T. Nishida [Ed], Conversational informatics: An engineering approach, 307–330, John Wiley & Sons, 2007.

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