Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 13–20
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Utterance-Unit Annotation for the JSL Dialogue Corpus: Toward a Multimodal Approach to Corpus Linguistics

Mayumi Bono [1&2], Rui Sakaida [1], Tomohiro Okada [2], Yusuke Miyao [3]
[1] National Institute of Informatics, [2] SOKENDAI (The Graduate University for Advanced Studies),
[3] The University of Tokyo
[1&2], 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JAPAN
[3], 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN
{bono, lui, tokada-deaf}@nii.ac.jp, yusuke@is.s.u-tokyo.ac.jp

Abstract

This paper describes a method for annotating the Japanese Sign Language (JSL) dialogue corpus. We developed a way to identify interactional boundaries and define a 'utterance unit' in sign language using various multimodal features accompanying signing. The utterance unit is an original concept for segmenting and annotating sign language dialogue referring to signer's native sense from the perspectives of Conversation Analysis (CA) and Interaction Studies. First of all, we postulated that we should identify a fundamental concept of interaction-specific unit for understanding interactional mechanisms, such as turn-taking (Sacks *et al.* 1974), in sign-language social interactions. Obviously, it does should not relying on a spoken language writing system for storing signings in corpora and making translations. We believe that there are two kinds of possible applications for utterance units: one is to develop corpus linguistics research for both signed and spoken corpora; the other is to build an informatics system that includes, but is not limited to, a machine translation system for sign languages.

Keywords: utterance unit, annotation, sign language dialogue

1. Introduction

This paper describes a method for annotating the Japanese Sign Language (JSL) dialogue corpus (Bono *et al*., 2014)[1]. Some linguists, including Deaf researchers who are interested in collecting sign language dialogue, began collecting data in April 2011. When we started, the general purpose of the project was to increase awareness of sign language as a distinct language in Japan. However, the academic aspects of the study recently became clear through interdisciplinary collaboration with engineering researchers, *i.e*., for natural language processing and image processing. In this paper, we introduce a preliminary result of our annotation process and annotated data, while explaining the concept of a 'utterance unit.' We anticipate that this concept will serve as a theoretical benchmark for promoting interdisciplinary research using spontaneous dialogue data in the corpus linguistics of sign languages.

2. Research Question and Background

In this study, we sought to find a way to identify interactional boundaries in sign languages and defined an utterance unit using various multimodal features accompanying signing.

2.1 Utterance Unit

The concept of utterance unit was already provided to segmenting and annotating spontaneous Japanese dialogues (Den *et al.* 2010; Maruyama *et al.,* in print). They propose a way of annotating utterance unit in two levels by emerging four linguistic and phonetic schemes, inter-pausal units, intonation-units, clause-units and pragmatic units.

In this paper, we define the concept of utterance unit for segmenting and annotating JSL dialogue data. We utilize JSL signer's native sense which is related to not only grammatical features but also multimodal features, such as mouth movements, non-manual movements, and gaze directions, to identify utterance unit. The method is based on classic observations in a research field of Conversation Analysis (CA) and Interaction Studies for spoken social interactions.

2.2 Sentence Unit

The previous studies on sign language linguistic have been focus on 'sentence unit' from the perspective of traditional linguistics. Crasborn (2007) introduces the workshop organized by his colleague and himself, which focuses on how to recognize a sentence in sign languages. He concludes that "we need to be alert to the risk of letting translations in another language influence our segmentation of signed language discourse, and keep our minds open for possible constructions that are modality specific" (Crasborn, 2007: 108).

Obviously, it should not rely on the writing system of spoken languages, because there is a risk of detecting an interactional chunk as a candidate of utterance unit using grammatical boundary of translated texts (e.g. JSL to Japanese). As widely known, there are some functional and grammatical utterance-final particles in Japanese, such as *ne* (ね), *yo* (よ), *yone* (よね) etc., they are possibly a signal of identifying interactional boundary. On the other hands, there is no functional and grammatical manual signs in JSL. In case of sign languages, these kinds of utterance final elements are spread in multimodal way, such as facial expressions and body postures.

2.3 Turn Constructional Units (TCUs) in CA

First of all, we had to introduce a classic concept of interaction-specific unit for understanding interactional mechanisms, such as turn-taking (Sacks et al., 1974). Conversation analysis (CA) is a sociological approach to the study of social interactions that applies the concept of turn constructional units (TCUs) (Sacks *et al.*, 1974) as fundamental building chunks of 'turns' in spoken

[1] Bono *et al*. (2014) introduces JSL colloquial corpus composed by dialogue part and lexicon part. Because we treat only dialogue part in this paper, we call it JSL dialogue corpus.

interactions, composed of utterances, clauses, phrases, and single words. CA research indicates that participants can anticipate TCUs and possible completion points of the ongoing turn using grammatical, prosodic, and pragmatic features of turn endings. Consequently, the turns in an interaction are exchanged smoothly among participants without difficulty.

Signers also naturally identify the boundaries of an utterance in social interaction, namely TCUs, to exchange turns visually. Signers probably recognize visual signals that are related to the grammatical, prosodic, and pragmatic completion points of turns. The concepts of TCUs and utterance units are similar. Here, we try to define an utterance unit in sign languages that aligns with the theoretical background of TCUs.

2.4 Applications

After identifying utterance units, we believe that they will have two applications: one is to develop corpus linguistics research for signed and spoken corpora; the other is to build an informatics system that includes, but is not limited to, a machine translation system for sign languages.

With regard to the former application, we anticipate that the research target of sign language studies will change drastically from example-based data to naturally occurring data, to study not only the grammatical aspects but also the social aspects of sign language interactions, such as turn-taking systems (Sacks *et al.*, 1974) and repair sequences (Schegloff *et al*., 1977) from the perspective of CA.

With regard to the latter application, we anticipate technical and theoretical breakthroughs in data collection and data storing using informatics technology, such as processing natural language and images. To recognize small hand and body movements in sign languages using image processing techniques (e.g., OpenPose), we will need to redesign the settings for data collection, lighting, frame rate, etc. If we want to translate sign language dialogue into spoken and written languages using deep learning or artificial intelligence technology, we will need to build a shared corpus to develop these systems.

The basic concept of the utterance unit is simple. However, we believe that it is a fundamental issue for developing sign language studies by combining research issues in linguistics and informatics.

3. Data

We collected JSL dialogues from 2012 to 2016. We have collected dialogues in 7 of the 47 Japanese prefectures (Table 1).

3.1 The first stage of data collection

As the first stage of data collection, we recorded videos of 40 deaf subjects in Gunma and Nara Prefectures (yellow in Fig. 1) from May to July 2012. Each prefecture has one school for the deaf. We obtained data from an age-balanced sample of individuals aged 30–70 years in each prefecture, and each age group was divided into same-sex pairs. Our participants from Gunma and Nara were in their 30s, 40s, 50s, 60s, and 70s, and included both male and female pairs.

3.2 The methods used for collecting dialogues

We used three methods to collect data: *interviews*, in which field workers and the assistants of native signers living in the same area who knew the procedures in advance asked the participants about their language, life, environment, etc. (for introductory purposes only, not open access); *animation narrative (AniN)*, in which one participant had memorized the story "Canary Row" and explained it to the other participant; and *lexical elicitation*, in which participants showed the corresponding signs for 100 slides of pictures and texts shown on a monitor, which is called JSL lexicon corpus (not included in this paper).

We collected pre-formed, lexical-level signing produced in a single-narrative setting and in spontaneous, utterance-level signing in a dialogue setting. In the single-narrative setting, we tried to detect enriched, deaf-specific signings using a theme for the narrative (*i.e*., folklore) and stimuli (pictures, images, etc.) to elicit signing at a lexical level. In a dialogue setting, we used video material to evoke a depictive signing (*i.e*., constructed action; Cormier, 2013) narrative task. We did not prepare a script for signing in advance. Consequently, the boundaries of the utterances were free, and were determined by participants who organized a turn-taking system in dialogue.

3.3 The amount of data

In the second stage of data collection, we collected data in Nagasaki, Fukuoka, Toyama, Ishikawa, and Ibaragi Prefectures, from 2014 to 2016 (green in Fig. 1). In this collection, we added two more dialogue tasks: '*my curry recipe (Cur)*' and '*proud of my country (Pro)*.'

Figure 1: Prefectures where dialogues were collected.

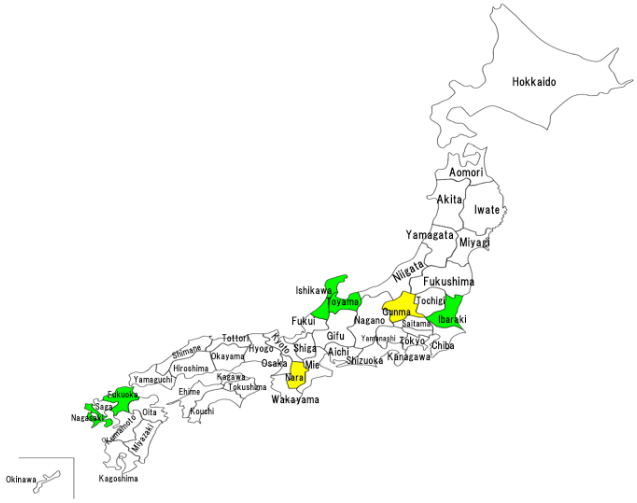

Table 1: Fundamental information of the dataset (collection year, number of dialogues, gender, and age range for each prefecture)

Prefecture	***Year***	***No. of Dialogue***	***Gender***	***Age Range***
Gunma	2012	10	M:10, W:10	30s-70s
Nara	2012	10	M:10, W:10	30s-70s
Nagasaki	2014	8	M:8, W:8	30s-70s
Fukuoka	2015	8	M:8, W:8	40s-80s
Toyama	2015	8	M:8, W:8	30s-70s
Ishikawa	2015	7	M:14, W:0	20s-80s
Ibaragi	2016	9	M:8, W:10	30s-70s
Total		60	120	20s-80s

Table 2: The percentage of video clips started putting basic annotations (word glosses and/or utterance unit glosses)

Prefecture	*AniN*	*Cur*	*Pro*	*Total*
Gunma	3/10			3/10
Nara	0/10			0/10
Nagasaki	8/8	8/8	4/8	20/24
Fukuoka	8/8	8/8	8/8	24/24
Toyama	8/8	8/8	4/8	20/24
Ishikawa	7/7	7/7	4/7	18/21
Ibaragi	0/9	0/9	0/9	0/27
Total	34/60 (56%)	31/40 (77%)	20/40 (50%)	85/140 (60%)

The total number of participants is 120 in 60 dialogues (See Table 1). The total recording time in corpus is 40 hours 52 minutes and 28 seconds. In the case in which we narrow down only dialogue tasks, AniN, Cur, and Pro, the total recording time is 15 hours 42 minutes and 59 seconds. As you can see in Table 2, the total number of video clips is 140. We have started putting basic annotations, word glosses and/or utterance unit glosses, to 85 files (60%). Actually, the annotated number of tokens, nearly equal to word gloss, is 27,371, November 26, 2019. Three independent video clips, collected camera A, B and C, were synchronized using Final Cut Pro. The original combined-angles image includes the interlocutor's back recorded by cameras B and C; there also is dead space, shown in black in Fig. 2. Cropped combined-angles images do not include the interlocutor's back and there is no dead space. The video images from all camera angles were enlarged to facilitate detailed analysis.2

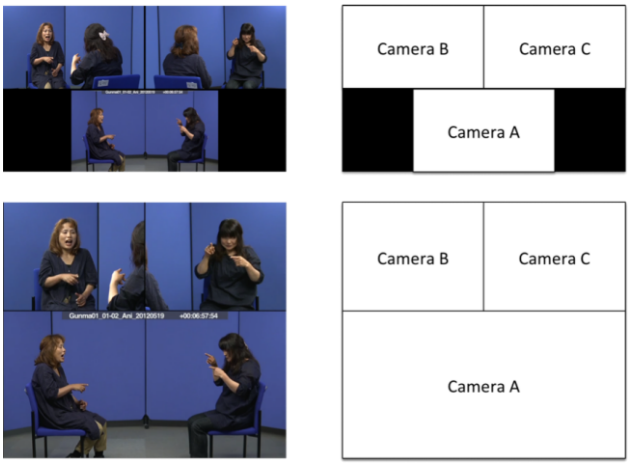

Figure 2: Image of two versions of the three camera angles: original (top), cropped (bottom).

4. Utterance Unit Annotation

We set two annotation levels: individual and integrated levels. The individual level is composed of four tiers: word gloss, mouth movement, non-manual movement (NMM), and gaze tiers. If annotators find features that can be used to a define utterance unit, such as the narrator's nodding behavior at turn-endings or a gaze shift from the signing space to the interlocutor, they classified them into each tier. Note that the annotated information does not include everything that happened in a dialogue; annotators tagged only information related to the grammatical, prosodic, or pragmatic features of turn endings.

At the integration level, all information annotated at individual levels is combined to identify utterance units. Figure 3 presents an example of the tier structure. 'NS_11_SH_40F' is the participant's information, which in this case means a female in her 40s who comes from the southern part (SH) of Nagasaki (NS), participant ID 11. Each tier has the participant's information to avoid confusing the annotated data among annotators. The tier names are placed after the participant information.

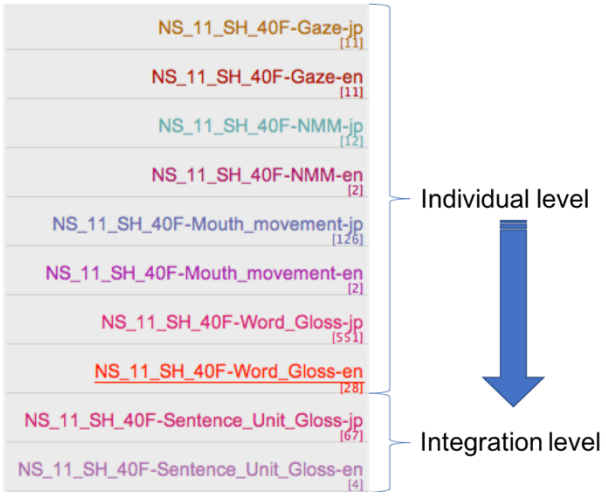

Figure 3: An example of the tier structure.

4.1 Individual Level

As mentioned above, the individual level is composed of four tiers. However, only the word gloss tier is mandatory; the others are optional.

4.1.1 Word Gloss Tier (mandatory)

Because JSL still does not have a digital dictionary or ID gloss system, such as *SignBank* and *Global Sign Bank*, the annotators placed the Japanese and English meanings of the signed word in the word gloss tier directly.

Before annotating the word gloss, all annotators learned the concept of the gesture unit (GU) proposed by Kendon (1970, 2004) to identify the start and end points of signed words.

One of our original plans was to establish a physical and hand movement unit smaller than the word gloss (Bono *et al.*, 2014). We applied the concept of the GU to annotate the beginning and end points of signed words. The GU is the interval between successive rests of the limbs, rest positions, or home positions. A GU consists of one or several gesture phrases. A gesture phrase is what we

[2] For more information about the JSL Dialogue Corpus, http://research.nii.ac.jp/jsl-corpus/research/data/manual/manual.html

intuitively call a 'gesture,' which consists of up to five phases: preparation, stroke, retraction, and pre- and post-stroke hold phases. We used the preparation, stroke, and retraction phases to identify the start and end points of a word gloss.

Figure 4 presents an image of a word unit for word gloss annotation. When several sign words form one utterance, such as in the lower image, the retraction phase is replaced by the preparation phase of the next signed word. In utterance unit annotation, annotators did not include the

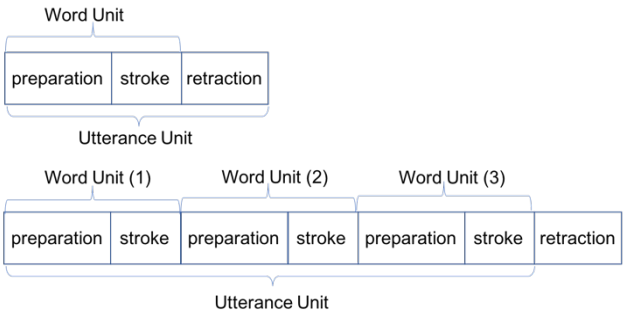

Figure 4: An image of word units for word gloss annotation.

GU phases, although they learned about this concept before annotating the word gloss.

Annotators write the meanings of signed words in capital letters on ELAN. If one signed word has a meaning that is a combination of two or more spoken words, such as TRY HARD (Excerpt 1, Section 5.1), it is connected by '.', *e.g.*, 'TRY.HARD', to indicate that it is one word in sign. If it is repeated several times, the number of repeats follows, *e.g.*, 'TRY.HARD (rep3)'.

4.1.2 Mouth Movement Tier (optional)

As in sign linguistics, there are two kinds of mouth movement in sign languages: mouth gestures and mouthing. A mouth gesture has a grammatical function, such as being adverbial to hand signings. Mouthing refers to shapes and movements that originate from spoken languages and are used while signing. Annotators classify these mouth movements by noting 'mg:' for mouth gesture and 'm:' for mouthing at the beginning of every annotation. When signers use mouth movements at the same time as signing, annotators place the following hand signing information in parentheses, *e.g.*, 'TRY.HARD(rep3) (m: ga-m-ba-ri-ma-su [try hard])' (Excerpt 1).

As mentioned above, we do not annotate all mouth movements, but only those related to utterance units, *i.e.*, the beginning or end of turns.

4.1.3 NMM Tier (optional)

We included a tier for non-manual movement (NMM) for grammatical and ungrammatical elements made by the body. For instance, a signer who has the role of a narrator may use nodding to show utterance endings. We would classify 'nod' into this tier.

4.1.4 Gaze Tier (optional)

Kendon (1967) observed the systematic mechanism of gaze direction at the ends of turns, including TCU endings and transition relevance places (TRPs) from the perspective of CA, from a psychological perspective. A shifting gaze can be crucial for identifying an utterance boundary, such as when signers shift their attention from the signing space to the interlocutor at the beginning or end of an utterance to confirm the interlocutor's understanding of the narrative. We would classify gaze directions into this tier.

4.2 Integration Level

Glosses at the individual level are combined at the integration level, which is the utterance unit. Annotators make a general judgement of the start and end of an utterance using information from individual levels annotated in advance.

All tiers in both levels are annotated by two native Deaf signers and one CODA (Children of Deaf Adults). Currently, these three annotators annotated five dialogues

Data ID	Gender of pair	Age	Task	Prefecture	Length of dialogue	(1) No. of Word Unit Gloss	(2) No. of Utterance Unit Gloss	Words in Utterance (1)/(2)
Data 1	Male	60's	*Cur*	Toyama	0:09:44	499	102	4.89
						624	119	5.24
Data 2	Female	60's	*Cur*	Toyama	0:08:04	304	38	8.00
						483	64	7.55
Data 3	Female	40's	*Cur*	Toyama	0:07:09	358	67	5.34
						490	70	7.00
Data 4	Female	40's	*AniN*	Toyama	0:10:04	896	109	8.22
						258	70	3.69
Data 5	Female	40's	*AniN*	Nagasaki	0:05:55	551	67	8.22
						57	33	1.73
Total	M:1; F:4	40's; 60's	*Cur*:3; *AniN*:2	Toyama:4, Nagasaki: 1	0:40:56	4,520	739	

Table 2: Results of Annotations.

on ELAN as a first test. As you can see in Table 2, total length of targeted five dialogues is almost 41 min. The average of tokens per min. is about 113[3].
Furthermore, table 2 shows the number of words gloss, the number of utterance unit gloss, and words per utterance.
Data 1, 2, and 3 are dialogues conducted the task of my curry recipe, and data 4 and 5 are dialogues conducted the task of animation narrative (Canary Row). There is a difference of the frequencies of utterance unit gloss between these tasks. In curry recipe, the number of words in utterance between participants in dialogues are balance such as 4.89 and 5.24 in data 1, on the other hand, in animation narrative, those are unbalanced, such as 8.22 and 1.73 in data 5. In Animation narrative task, there is the tendency that the participant who watched movie clip in advance holds turns and have multiple TCU in a turn, and the intelocuter gives small number of words to narrator in short reactions, such as, *uhn hm, I see* in English.

5. Analysis

In the following analyses, we present three excerpts analyzed using the CA framework to clarify how we integrate the features in tiers at the individual level to identify utterance units.

5.1 Excerpt 1: TRPs with Mouthing

In excerpt 1, signer TY_12 (lower tiers in ELAN of excerpt1, Figure 5) says, "*You should cook a delicious meal for your husband*. (HUSBAND/ FOR/ DELICIOUS/ MAKE/ GIVE (m: a-ge-te [give])". Signer TY_11 answers by mouthing and signing, "*Yes, I'll do my best.* ((m: ha-i [yes])/ TRY.HARD (rep3) (m: ga-m-ba-ri-ma-su [try hard]))".

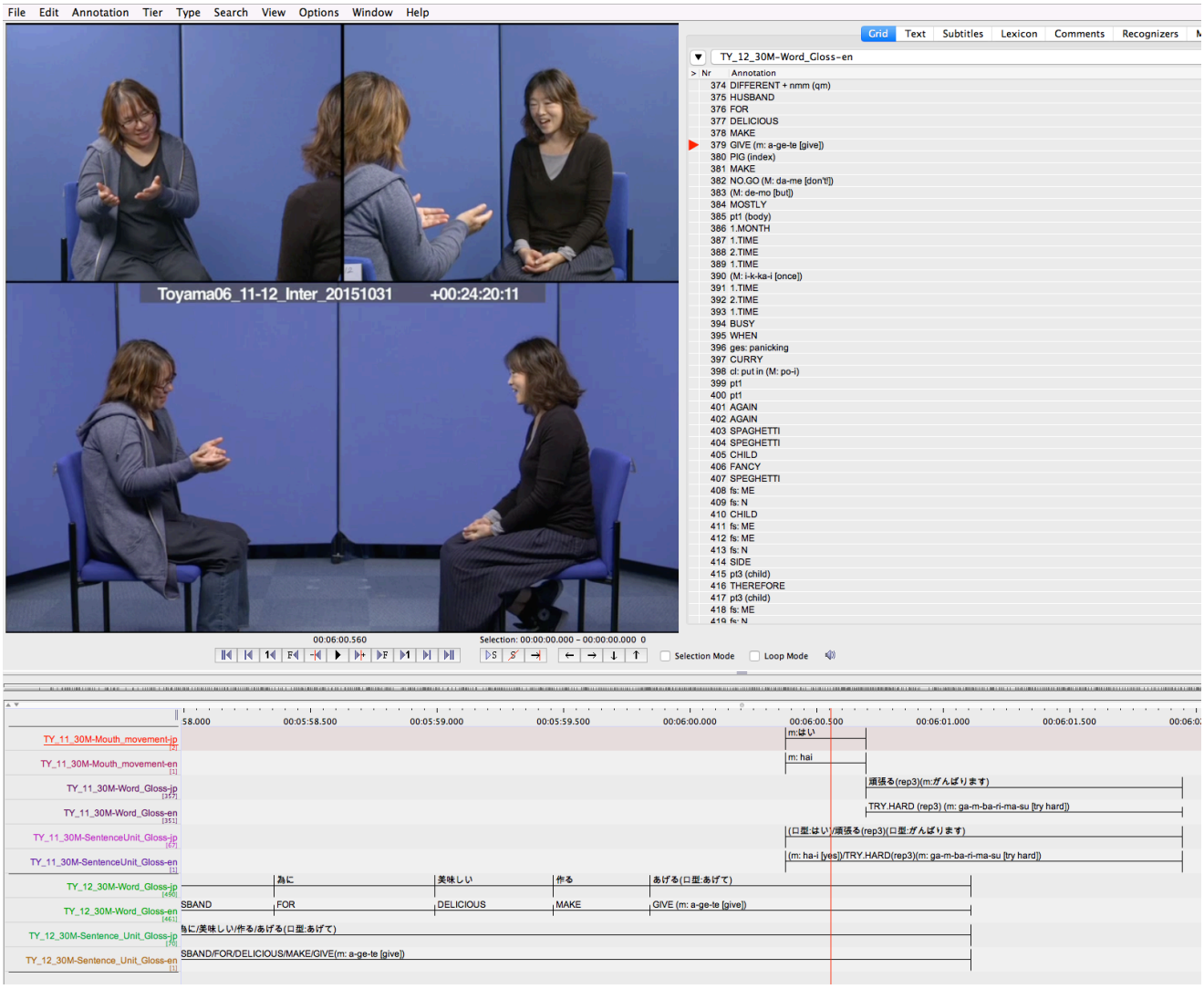

Figure 5: ELAN annotation of Excerpt 1, Transition relevance place with mouthing.

[3] Reviewer #1 pointed out to us that the average of tokens per min in a dialogue setting is about 120 in DGS (German Sign Language). There is a similar tendency in our corpus.

Interestingly, TY_11 produces mouthing before hand signing to answer TY_12's recommendation. Specifically, during the middle of the final sign /GIVE/, TY_12 starts her answer by mouthing 'ha-i [yes]'. /GIVE/ is a subsidiary verb in Japanese and is also an agreement verb in JSL. TY_12 moves both of her hands to the right, where the semantic meaning "TY_11's husband" was given in advance (Figure 5) (Liddell, 2012). Before signing /GIVE/, TY_12's utterance is almost grammatically and semantically completed. Consequently, within /GIVE/, specifically close to the end of the stroke for /GIVE/, is the earliest sequential position for TY_11 to give a response. And TY_11 gives a first response, not by signing, but by mouthing 'm: ha-i [yes]'.

In this segment, their utterances overlap. We assume that this is a typical case of transition relevance place (TRP) in sign language dialogues. It implies that we should include mouthing when discussing the utterance units.

5.2 Excerpt 2: Narrative and Role-shift with Gaze

Next, we discuss utterance completion and the narrator's gaze behavior. In excerpt 2, signer NS_11 (upper tiers in the ELAN of excerpt 2, Figure 6) is telling a story about the animation she has watched, called 'Canary Row' which is an animation clip used in Gesture Studies (e.g. McNeill, 1992).

She describes a famous first scene of it by producing multiple utterances: (1) "*A chick is swinging inside a bird cage*. (cl:human:un(stop)/ cl:human:chick:having a swing(stop)/ cl:sphere(circle)/ cl:un (something that swings like a swing in the sphere)/ cl:the shape of a cage/)"; (2) "*The cute chick is playing on the swing.* (pt3(rep)/ CUTE/ cl:human:chick: (ges:flaps the wings)/ cl:human:chick: (ges: enjoying playing on a swing)"; and (3) "*A cat finds the chick, and climbs something like a pillar quickly*. (pt1 (meaning: pt (cat))/ CAT/ cl:human: cat: (ges: looks around, notices something and claps his hands)/ cl:human:cat:climbs something like a pillar(stop)/ cl:explanation of something like a pillar standing upright/cl:human:cat (ges:looks around quickly) /cl: human: cat climbs the pillar quickly /cl:climbs/ cl:climbs+NMM)".

As we can see, she uses lexical expressions only for /CUTE/ and /CAT/ in this part. Moreover, these lexical signs are accompanied by mouthing. Other expressions are depicting signs (cl) and gestures (ges) without mouthing. This style of signing is very familiar in sign language narrative talk.

We focus on the narrator's gaze behavior at the boundary of each utterance. In each ending of all the utterances, the narrator (NS_11) looks at the interlocutor (NS_12). Furthermore, the narrator gives a nod at the end of utterances (1) and (3) and the interlocutor gives a response, such as /UNDERSTAND/ at these points.

The analysis of excerpt 2 revealed that the gaze directions and head nods accompanying the narrative provide clues for the interlocutors to identify utterance units. Moreover, the integration of these clues shows the utterance boundary more strongly.

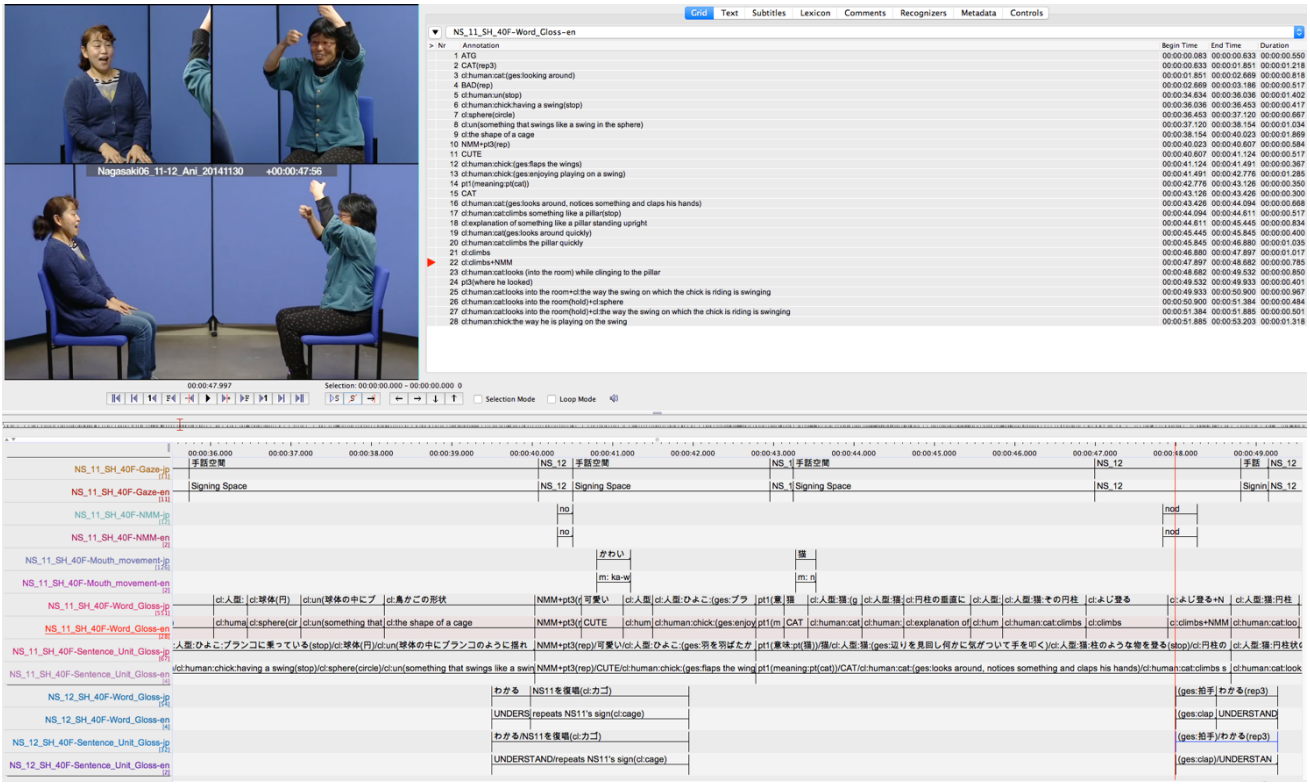

Figure 6: ELAN annotation of Excerpt 2, Narrative and role-shift with gaze.

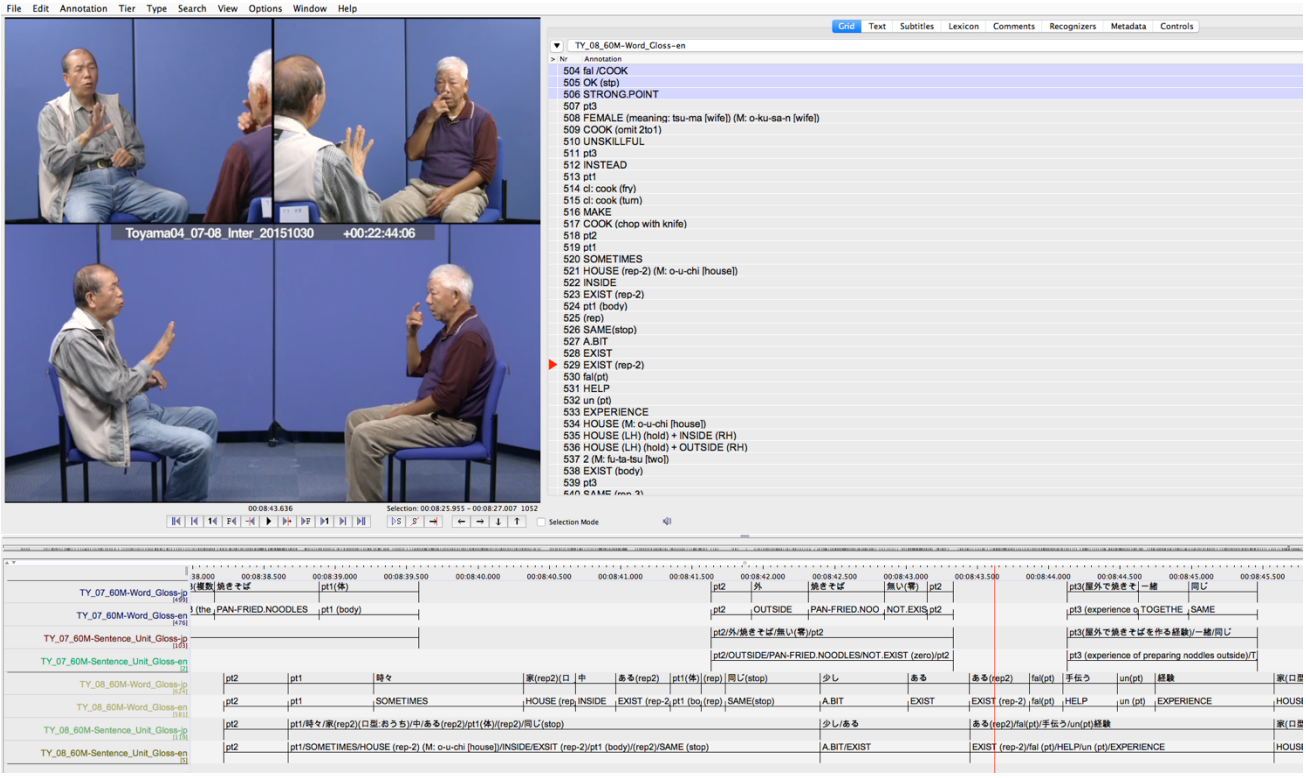

Figure 7: ELAN annotation of Excerpt 3, Other-initiated repair with overlap.

5.3 Excerpt 3: Other-initiated Repair with Overlap

Finally, excerpt 3 is an example in which they exchange their utterances orienting to turn-taking system, not a narrative or one-way signing. Excerpts 2 and 3 appear similar at the point where the interlocutor's actions – responses in excerpt 2 and confirmation questions in excerpt 3 – overlapped with the current signer's signing and are retroactively defined as utterance units.

In excerpt 3, TY_08 starts to explain his experience cooking pan-fried noodles (*yakisoba*) by producing sequential multiple utterances, (1) "*I'll make it in my house*. (pt1/ SOMETIMES/ HOUSE (rep-2) (M: o-u-chi [house])/ INSIDE/ EXIST (rep-2)/ pt1 (body)/ (rep2)/ SAME (stop))", (2) "*Sometimes…* (A.BIT/EXIST)", and (3) "*Yeah, I helped cooking*. (EXIST (rep-2)/ fal (pt)/ HELP/ un (pt)/ EXPERIENCE)".

However, these three utterances are not connected like the narrative in excerpt 2. From TY_07's questions, we can see how TY_08's multiple sequential utterances are connected. That is, TY_07 asks TY_08 a question to display his understanding (Sacks, 1992). TY_08 answers him as soon as possible, as in the utterance (2) mentioned above.

To obtain more detail, although TU_08 continues his turn with the utterance (1), TY_07 asks him, "*Don't you make them outside* (like camping)? (pt2/ OUTSIDE/ PAN-FRIED.NOODLES/ NOT.EXIST (zero)/ pt2)" during the utterance's final particles, pt1(rep), which is a sandwich construction with pt1 in the utterance-onset, which is a TRP. Then, TY_08 gives an answer by connecting his utterance with his previous utterance, as in utterance (2), "*Sometimes…*". That is the earliest place for him to give a response.

The exchanges in excerpt 3 are related to the concept of other-initiated repair sequence (Schegloff, 1977). There are four techniques for others to initiate repair: open class forms, category-specific interrogatives, repeats of the trouble-source turn, and candidate understandings (Sidnell & Stivers, 2013). A TY_07's question overlapping an utterance (1) is used as pt2 (pointing at the interlocutor) at the onset and offset of an utterance, which makes it clear that there is a something trouble for TY_07 in TY_08's utterance, who was pointed out by pt2.

Consequently, the annotators segmented TY_08's signing into three parts: utterances (1), (2), and (3). Utterance (1) is a description of his experience; utterance (2) is an answer to TY_07's question; and utterance (3) is an elaboration of his own answer in utterance (2).

From the analysis of excerpt 2, we found that not only annotated multimodal features in tiers at the individual level, but also the sequential structure of dialogues, are clues used to identify utterance units.

6. Conclusions

This paper describes an annotation method for the Japanese Sign Language (JSL) dialogue corpus (Bono *et al.*, 2014) by defining the concept of an utterance unit. By analyzing three excerpts, we showed how complicated it is to identify utterance units using a combination of signing and various other features. However, annotators who are all native signers (Deaf and Coda) with a native understanding of JSL

used multimodal features to identify the utterance units. We found that it was very difficult to establish a standard criterion for finding features among annotators.

The utterance is a fundamental unit in languages. It is obvious that we cannot rely on the written system of spoken languages. Defining the utterance unit in sign languages will have useful applications, such as setting a fundamental unit for storing data in sign language corpora, and for manual or machine translation using advanced technologies.

7. Bibliographical References

Bono, M., Kikuchi, K., Cibulka, M. and Osugi, Y. (2014). Colloquial corpus of Japanese Sign Language: A design of language resources for observing sign language conversations. Proc. of the ninth International Conference on Language Resources and Evaluation Conference, pp.1898-1904.

Crasborn, O. (2007). How to recognise a sentence when you see one. *Sign Language & Linguistics* **10-2**, pp.103-111.

Den, Y., Koiso, H., Maruyama, T. and Yoshida, N. (2010). Two-level Annotation of Utterance-units in Japanese Dialogs: An Empirically Emerged Scheme. Proc. of Seventh International Conference on Language Resources and Evaluation, pp.17-23.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, **26**, pp.22-63.

Kendon, A. (1970). Movement coordination in social interaction. *Acta Psychologica*, **32**, pp.1-25.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Liddell, S. (2012). *Gesture, Grammar and Meaning in American Sign Language*. Cambridge University Press.

Maruyama, T., Den, Y. and Koiso, H. (in print). Design and annotation of two-level utterance units in Japanese. In Izre'el, S. *et al*. (eds) *Search of Basic Units of Spoken Language: A corpus-driven approach.* John Benjamins.

McNeill, D. (1992). *Hand and mind*. Chicago, IL: University of Chicago Press.

Sacks, H., Schegloff, E. A. and Jefferson, G. (1974). A Simplest Systematics for the Organisation of Turn-Taking for Conversation, *Language*, **50**, pp.696–735.

Sacks, H. (1992). *Lectures on Conversation, Volumes I and II,* Edited by G. Jefferson with Introduction by E.A. Schegloff, Blackwell, Oxford.

Schegloff, E. A., Jefferson, G. and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation, *Language*, **53**(2), pp. 361-382.

Sidnell, J. and Stivers, T. (2013). *The handbook of conversation analysis*. Wiley-Blackwell.

8. Appendix

HUSBAND	A sign is indicated in upper case.
NOT.EXIST (zero)	When there are more than two expressions for what is being signed, the expression selected is indicated in parentheses ( ).
cl:the shape of a cage	Classifiers or depicting signs are indicated in lower case.
(m: a-ge-te [give])	Mouthing. Hyphens (-) are used to delimit each kana syllable. The translation of the mouthing is added within brackets [ ].
pt1	Pointing to speaker him/herself
pt2	Pointing to hearer
pt3	Pointing to neither speaker nor hearer
pt1(body)	The object pointed to is indicated in lower case within parentheses ( ).
(ges:flaps the wings)	Gestures
NMM	Non manual markers
(rep3)	Reduplications. When the number of iterations is known, it is indicated as "(rep2)" for two iterations, "(rep3)" for three and so forth.
(stop)	A cut-off or truncation
un	Unclear hand movements
fal	Signing errors
(meaning: pt (cat))	The meaning of a sign in the conversational context is sometimes described.