Analysis of Child-Caregiver Interactions for Developing a Caregiver Spoken Dialogue System

Sanae Yamashita[1], Shota Mochizuki[1], Yuko Kuma[2], Ray Sakai[3], Ayaka Sasaki[4], Ryuichiro Higashinaka[1]

[1]Nagoya University, Japan [2]Mamoru Co., Ltd., Japan
[3]Logical Fabrics, Inc., Japan [4]RainbowWanderlust Co., Ltd., Japan

Correspondence: yamashita.sanae.w7@s.mail.nagoya-u.ac.jp

Abstract

We aim to develop a caregiver spoken dialogue system for remote childcare services. As a first step toward this goal, this study investigates how interactions occur between children and caregivers. We collected Japanese child–caregiver dialogue data through a remote childcare service in which participants engaged in activities such as introductions, quizzes, and free conversations. The collected data were analyzed and compared with existing child–caregiver dialogue data from both acoustic and linguistic perspectives. The results showed that, acoustically, child–caregiver dialogues contained fewer overlapping utterances than adult dialogues. Linguistically, the distribution and transitions of utterance intentions differed across dialogue parts, reflecting the diverse structures of each activity. These findings provide useful insights for building future caregiver spoken dialogue systems, suggesting that a turn-based interaction structure may be sufficient and that dialogue control should be adapted to each part of the dialogue.

1 Introduction

The declining birthrate and the shortage of childcare professionals have increased the demand for systems that can support children remotely. In addition, recent social changes have heightened the importance of remote communication, drawing attention to non-face-to-face childcare support and home-based childcare dialogues. Under these circumstances, spoken dialogue systems have the potential to serve as one means of supporting children's development while addressing the shortage of human caregivers.

Against this background, spoken dialogue systems for children are being actively researched, adding to a growing trend of dialogue research involving children (Rakhymbayeva et al., 2021; Xu et al., 2022; de Haas et al., 2022; Kim et al., 2022).

RainbowWanderlust Co., Ltd. operates a remote childcare service called NannyME[1], in which caregivers can interact and communicate with children through spoken interaction. Although the current service is operated by human caregivers, its operational cost remains high. Considering the increasing demand for childcare support, such services should preferably be partially automated.

This study takes the first step toward developing a caregiver spoken dialogue system for remote childcare services by investigating how interactions occur between children and caregivers. Specifically, we collected a Japanese child–caregiver dialogue dataset using our service and analyzed its characteristics. In particular, we comparatively analyzed it with existing child–caregiver dialogue corpora in terms of turn-taking, backchannels, and utterance intentions, aiming to obtain insights that can inform future system design.

2 Related Work

This section reviews prior studies on spoken dialogue corpora involving children and adults, analyses of dialogues that include children, and spoken dialogue processing designed for interactions with children.

2.1 Construction of Spoken Dialogue Corpora

Dialogue corpora involving children and adults have been developed mainly by recording natural parent–child interactions. A representative example is the CHILDES corpus[2] (Macwhinney, 2000), which contains recordings and transcriptions of parent–child dialogues in multiple languages, including Japanese. The CHICA corpus (Goumri et al., 2024) includes recordings and transcriptions of parent–child interactions annotated with dialogue phenomena such as utterance intentions, repairs,

[1] https://nannyme.love/
[2] https://talkbank.org/childes/

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 357–368
February 26–March 1, 2026.

and backchannels. Similarly, ChiCo (Bodur et al., 2021) provides recordings and transcriptions of parent–child dialogues in various languages, along with acoustic features and dialogue act labels.

Although corpora specifically focused on Japanese are limited, CEJC-Child (Koiso et al., 2025) was recently constructed by recording and transcribing natural parent–child interactions in daily life, covering diverse contexts such as play and mealtimes. R-JMICC (Saikachi et al., 2013) consists of recordings of scenes in which infants and their mothers play with picture books and toys. In the present study, we independently collect spoken dialogue data of child–caregiver interactions using our remote childcare service and analyze its characteristics.

2.2 Analysis of Spoken Dialogues with Children

Dialogues involving children have been studied primarily from the perspectives of turn-taking and backchannels.

Regarding turn-taking, comparative analyses of natural conversations between children and adults have reported that children's utterances contain fewer overlaps (Gallagher and Craig, 1982; Horváth and Krepsz, 2023). In addition, children respond more slowly than adults because they expend a greater cognitive load on language processing (Casillas et al., 2016). From the viewpoint of backchannels, Bodur et al. (2023) annotated backchannels in a parent–child dialogue corpus and showed that children tend to produce more content-responsive (specific) backchannels than their parents. Research has also focused on utterance intentions. Ninio et al. (1994) proposed the INCA-A label set, which classifies children's utterances in accordance with communicative intent based on English mother–child dialogues. Nikolaus et al. (2022) applied an utterance intention classification model derived from INCA-A to the CHILDES corpus and demonstrated that the variety of utterance intentions increases with age.

Following these studies, we analyze our data from the perspectives of turn-taking, backchannels, and utterance intentions.

2.3 Spoken Dialogue Systems for Children

In the educational domain, studies have implemented learning support tutors using speech recognition (Mostow et al., 2003; Ward et al., 2011) and explored robot-assisted vocabulary learning (Kory-Westlund and Breazeal, 2019). In the welfare domain, dialogue robots have been developed to support speech training (Esfandbod et al., 2023) and provide social assistance for children with autism spectrum disorders (Biagi et al., 2025).

Liu et al. (2022) proposed a backchannel prediction model that estimates adult backchannel timing from children's speech using machine learning, thereby enabling automated control of backchannel behavior during dialogue. Moreover, an automatic evaluation benchmark for child–caregiver dialogues (Liu and Fourtassi, 2024) has been created, providing a framework for objectively comparing speech recognition and response generation models.

Our research aims to develop a spoken dialogue system able to interact with children within our remote childcare service.

3 Method

We ultimately aim to develop a caregiver spoken dialogue system for remote childcare services. As a first step toward this goal, we seek to clarify how interactions occur between children and caregivers. The procedure is as follows. First, we collect dialogue data from our remote childcare service NannyME. Then, we compare the collected dialogues with existing Japanese spoken dialogue corpora that include interactions between adults (caregivers) and children. Specifically, we compare our data with parent–child dialogues included in CEJC-Child (Koiso et al., 2025) and CHILDES (Macwhinney, 2000). The comparison focuses on three aspects: turn-taking, backchannels, and utterance intentions.

3.1 Data Collection

To focus on our remote childcare service, we collect Japanese dialogue data that includes interactions characteristic of the service. Specifically, the service often involves casual conversation and quizzes. Based on this, we collect dialogues that start with an introduction part, proceed through a quiz part and a free conversation part related to the quiz answers, and end with a closing part. In each dialogue, a quiz part and a free conversation part are combined consecutively several times (three times in this study).

The introduction part takes place at the beginning of the dialogue. In this part, the caregiver greets the child and the guardian, after which the

Category	Abbreviation	Expressions
Responsive interjections	R	*hai, un, aa, ee, hun*
Expressive interjections	E	*a, e, o, hee, huun*
Lexical reactive expressions	L	*soo(-desu-ne) (I think so), naruhodo (really), tashikani (surely), ne (huuh)*
Assessments	A	*sugoi (great), omoshiroi (funny), kowai (terrible)*

Table 1: Categories and expressions of backchannels used in this study, based on the types of reactive tokens defined in (Den et al., 2011).

child and the caregiver share their personal attributes. They freely talk about their names, favorite foods, favorite animals, and similar topics. In the quiz part, the caregiver presents a quiz to the child. The quiz follows the "three-hint" format, where the caregiver prepares the correct answer and three related hints in advance, and discloses the hints one by one. After each hint, the caregiver asks the child to guess the answer. When the child gives the correct answer, or when the caregiver judges that the child cannot work it out, the caregiver reveals the correct answer. The free conversation part follows the quiz and allows for open dialogue. The caregiver starts the conversation on the basis of the quiz answer and develops it in accordance with the child's interests. In the closing part, the caregiver looks back on the dialogue and asks the child for their impressions of the session.

3.2 Comparison of Turn-Taking

For the comparison of turn-taking, we follow (Nguyen et al., 2023) and compare four measures: IPU (the duration of an inter-pausal unit, defined as a speech segment separated by a silence longer than 0.2 seconds), Pause (the duration between IPUs), Gap (the duration of silence between utterances by different speakers), and Overlap (the duration of overlapping speech between different speakers). These measures are calculated as cumulative durations per minute. IPU and Pause are computed separately for the child and the caregiver, while Gap and Overlap are calculated as a single value for each speaker pair.

3.3 Comparison of Backchannels

For comparing backchannels, the backchannels in the transcriptions are classified by category, and their occurrence frequency per actual speaking time is compared. The counting procedure follows the backchannel categories (R, E, L, and A) proposed by Den et al. (2011) and the backchannel expressions shown in Table 1. Specifically, using regular expressions, we count a match when a string in the text exactly matches one of the listed words or consists of repetitions of the same listed expression.

Regarding category R, Kawahara et al. (2016) reported that the function of response tokens differs depending on whether they are repeated once, twice, or three times or more. Therefore, in this study, we count them separately as distinct categories (R1, R2, R3) in accordance with the number of repetitions. Expressions such as "aa" or "ee" could be interpreted as E tokens, but in this study, they are counted as R tokens.

3.4 Comparison of Utterance Intentions

For comparing utterance intentions, each utterance is classified in accordance with the speaker's communicative intent, and the frequency and transitions of utterance intentions are analyzed.

The classification is based on a modified version of INCA-A. In the original INCA-A framework, 12 categories are defined, and both a preceding utterance (e.g., a question) and its corresponding response (e.g., an answer) belong to the same category. However, since separating questions and responses was considered more informative for analysis, we broke down these categories and used a total of 19 categories.

Utterance intentions are automatically classified using a large language model (LLM). Specifically, we use an LLM provided by OpenAI[3], where definitions of the 19 categories and example utterances are provided for the prompt (Fig. 4 in Appendix). Up to 50 utterances of dialogue history are given as input to the model, and the model outputs the most appropriate category name for each utterance. In accordance with the terms of use, comparisons are made only with the CHILDES corpus. In a preliminary experiment, one of the authors manually annotated 50 sampled utterances with category labels and compared them with the model's outputs. The results showed a Cohen's kappa value of 0.71, indicating relatively high consistency.

For analyzing utterance intention transitions, we visualize the transition probabilities between dia-

[3] https://platform.openai.com/docs/models/gpt-5-nano

Introduction	
Caregiver	Hello.
Child	Hello.
Caregiver	Ah, thank you.
Caregiver	Can you tell me your name?
Child	I'm [Name].
Caregiver	[Name], how old are you?
Child	Four years old.
Quiz	
Caregiver	Let's listen to the second hint.
Caregiver	It has a beard.
Caregiver	Oh, is it you, [Name]?
Child	A lion.
...	...
Caregiver	But there are many animals with beards, right? Goats have them too, don't they?
...	...
Caregiver	I say "meow."
Child	A cat?
Caregiver	A cat!
Caregiver	Correct!
Free conversation	
Caregiver	Have you ever touched a cat?
Child	I've seen one.
Caregiver	Oh, you've seen one? Where did you see it?
...	...
Child	They like milk and fish.
Caregiver	Uh-huh.
Caregiver	Uh-huh.
Caregiver	You know a lot!
Child	Because I read it in a book.
Caregiver	You read it in a book?
Caregiver	Wow, that's great! You're such a good learner.
Closing	
Caregiver	Ms. [Name] is getting sleepy. My eyes are like this.
Caregiver	Right?
Child	Yeah.
Caregiver	Getting sleepy, huh? My cheeks are drooping and my eyes are closing.
Caregiver	Oh, [Name], it's almost time to say goodbye.
Caregiver	Did you have fun?
Caregiver	I had fun too!

Table 2: Examples of dialogues from the collected data. The utterances were originally in Japanese and were translated by the authors.

logue acts as graphs and qualitatively examine the flow of interactions.

4 Collection of Remote Childcare Dialogue Data

We collected dialogues between children and caregivers through our remote childcare service, which enables communication via video calls. This data collection was approved by the ethics review board of our institution. All participants and their guardians provided informed consent for the results to be collected, analyzed, and published.

4.1 Dialogue Collection

We recruited children and caregivers as speakers from users of our service. All speakers were native speakers of Japanese, and we ensured the gender distribution was as balanced as possible. The recruited children were between four and six years old, an age range suitable for spoken communication with caregivers. The caregivers were those who provided childcare within our service. In total, the dataset included 27 children (10 boys and 17 girls) and 10 caregivers (4 males and 6 females). At the beginning of data collection, eight children were four years old, seven were five, and 12 were six.

We collected 50 approximately 30-minute videos of one-on-one dialogues between children and caregivers. The dialogues were conducted via video calls using Twilio Video[4] or Agora[5], and the speakers participated using a PC, tablet, or smartphone. Each child participated in up to two dialogue sessions, while the number of sessions per caregiver ranged from one to 18. Each child–caregiver pair appeared only once, meaning that all 50 dialogues involved unique combinations of children and caregivers.

Before recording, caregivers were instructed on the dialogue flow, which consisted of an introduction part, quiz part, free conversation part, and closing part. They were also advised to focus on spoken interaction and to avoid, as much as possible, using physical objects, gestures, or play activities.

All collected dialogues were manually transcribed. Examples of the collected dialogues are shown in Table 2.

4.2 Dialogue Statistics

Table 3 shows the statistics of our dataset, referred to as the Remote Childcare Dialogue Data (RCDD), along with those of the comparison corpora. For CEJC-Child, we used only the dialogues involving two speakers from the publicly available monitor version. For CHILDES, we used the two-speaker dialogues in the MiiPro subset[6] (Miyata, 2012), which provides both transcriptions and timestamps of utterances. Although all corpora consist of dialogues between a child and a caregiver, note that the caregivers in the RCDD are non-parental

[4] https://twilio.com/docs/video/

[5] https://www.agora.io

[6] https://talkbank.org/childes/access/Japanese/MiiPro.html

	Remote Childcare Dialogue Data (RCDD)		CEJC-Child (Koiso et al., 2025)		CHILDES (Macwhinney, 2000)	
	Child	Caregiver	Child	Caregiver	Child	Caregiver
Child age range	4–6 yrs		0–8 yrs		1–2 yrs	
Number of dialogues	50 dialogues		53 dialogues		70 dialogues	
Total dialogue duration	23 hrs (28 mins)		13 hrs (15 mins)		76 hrs (65 mins)	
Actual speaking time	4.5 hrs (5 mins)	13.2 hrs (16 mins)	3.0 hrs (3 mins)	5.1 hrs (6 mins)	16.9 hrs (14 mins)	32.6 hrs (28 mins)
Speaking ratio	19.4%	57.0%	22.5%	37.7%	22.1%	42.8%
Number of unique speakers	27	10	7	11	3	15

Table 3: Statistics of the Remote Childcare Dialogue Data compared with other Japanese child–caregiver dialogue corpora. Values in parentheses indicate averages per dialogue.

Dataset	IPU		Pause		Gap	Overlap
	Child	Caregiver	Child	Caregiver	–	–
Introduction	11.0 (1.5)	37.2 (3.4)	3.1 (2.4)	7.7 (1.4)	8.6 (1.1)	3.1 (0.6)
Quiz	3.7 (1.3)	15.2 (3.6)	1.3 (3.2)	22.0 (11.9)	7.7 (3.1)	1.2 (0.6)
Free conversation	8.3 (2.0)	17.3 (2.7)	3.5 (5.9)	9.9 (2.9)	10.0 (3.6)	2.0 (0.6)
Closing	0.7 (1.5)	2.5 (4.4)	0.2 (1.3)	0.6 (1.3)	0.5 (1.4)	0.3 (0.7)
CEJC-Child	12.6 (1.2)	21.7 (1.5)	3.8 (1.8)	11.9 (1.8)	13.0 (1.3)	1.3 (0.4)
CHILDES	13.3 (2.1)	25.5 (2.9)	3.3 (3.0)	9.3 (2.4)	8.1 (2.0)	3.0 (0.5)
Adult–adult	–	59.7	–	3.5	4.0	8.1

Table 4: Comparison of total durations (in seconds per minute) of IPUs, pauses, gaps, and overlaps. Values in parentheses indicate the average duration per instance. The reference values for adult–adult dialogue are taken from (Ohashi et al., 2025).

professionals, whereas those in CEJC-Child and CHILDES are the children's parents.

The age range of the children also differs across corpora. The RCDD includes preschool children aged 4–6 years, CEJC-Child covers infants to elementary-aged children (0–8 years), and CHILDES consists mainly of toddlers aged 1–2 years.

In terms of total dialogue duration, the RCDD contains less total dialogue duration than CHILDES but more than CEJC-Child. The average duration per dialogue is shortest in CEJC-Child, about twice as long in the RCDD, and roughly twice again in CHILDES. In every dataset, caregivers speak approximately two to three times longer than children, and this ratio is highest in the RCDD; the child–caregiver speaking ratios in each part were as follows: 19.6% vs. 63.7% in the introduction part, 6.2% vs. 46.9% in the quiz part, 27.8% vs. 40.6% in the free conversation part, and 8.6% vs. 86.2% in the closing part.

The numbers of unique speakers are 27 children and 10 caregivers in the RCDD, indicating greater variation among child speakers compared with the other corpora.

5 Analysis

We comparatively analyzed the RCDD, CEJC-Child, and CHILDES from three perspectives: turn-taking, backchannels, and utterance intentions.

5.1 Comparison of Turn-Taking

Table 4 shows the measures related to turn-taking.

For IPUs, child–caregiver dialogues as a whole tended to have shorter IPUs than adult–adult dialogues. This indicates that utterances were generally shorter and that silent intervals occurred more frequently in child–caregiver dialogues. Among the child–caregiver datasets, the caregiver's IPUs were longer in the introduction part of the RCDD. In contrast, the children's IPUs were generally short, particularly in the quiz and closing parts. This suggests that children spoke to some extent during the introduction and free conversation parts, but spoke less during the quiz and closing parts.

Regarding the average length of each IPU, caregivers generally had longer IPUs than children in all corpora, but the difference varied across datasets. In the RCDD, the difference between caregivers and children was larger in the introduction, quiz, and closing parts, and smaller in the free conversation part. In CEJC-Child, both speakers had short IPUs, while in CHILDES, IPUs were relatively long. However, in both CEJC-Child and CHILDES, the difference between children and caregivers was small.

For Pause and Gap, except for the closing part of the RCDD, child–caregiver dialogues showed

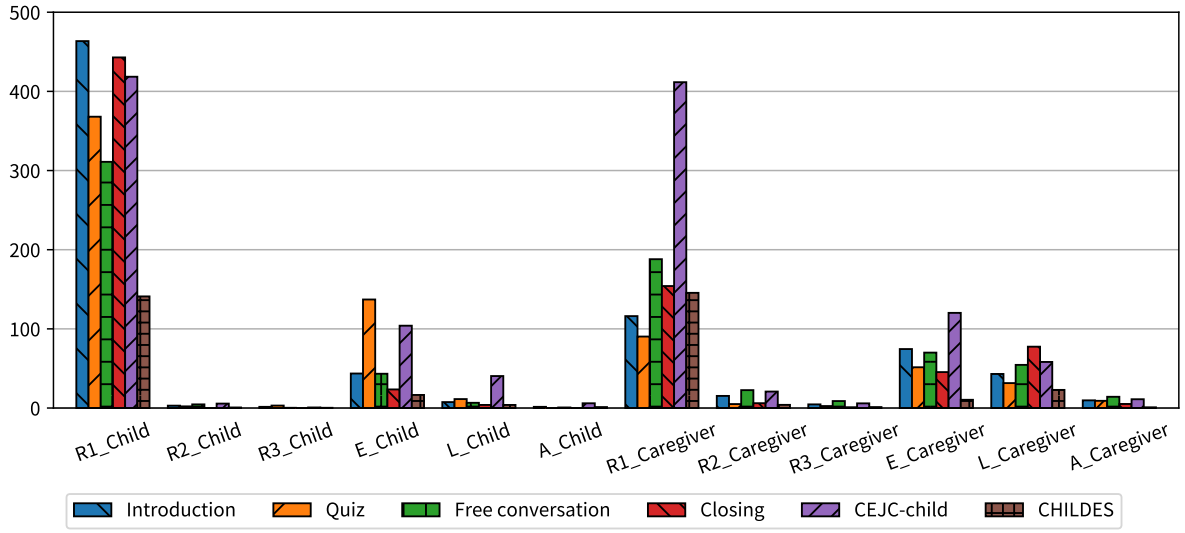

Figure 1: Frequency of backchannels produced by children and caregivers per hour of speaking time.

longer pauses and gaps than adult–adult dialogues. This indicates that, overall, the speaking pace was slower in child–caregiver dialogues than in adult–adult dialogues. Across the child–caregiver corpora, the total pause time of caregivers was longer than that of children, though the average duration per pause differed. Specifically, caregivers' pauses were generally about one to two seconds, but exceeded 10 seconds in the quiz part. In addition, in the quiz part, the average gap duration per occurrence was also relatively long (3.1 seconds), indicating that there were longer silent periods during the quiz.

For Overlap, child–caregiver dialogues generally showed shorter overlaps than adult–adult dialogues. The longest overlaps were observed in the introduction part of the RCDD and CHILDES, but even in those cases, the duration was only about three seconds per minute. This shows that the speech of children and caregivers rarely overlaps.

5.2 Results of Backchannel Comparison

Figure 1 shows the frequency of backchannels per hour of speaking time.

For children, R1 was found to be the most frequent category overall, particularly in the RCDD and CEJC-Child. In the RCDD, the frequency of R1 varied across dialogue parts, appearing relatively often in the introduction and closing parts. E tokens were observed more frequently in the quiz part, with "e" and "a" occurring often. These expressions are assumed to indicate confusion or hesitation in response to quiz hints, or that the child was thinking. R2, R3, and A tokens were rarely observed. The increased use of E tokens during the quiz part indicates that children use expressive interjections not only for feedback but also to externalize their thinking process, revealing a cognitive rather than purely reactive function of backchannels.

For caregivers, the RCDD showed a lower overall frequency of R1 than CEJC-Child. Within this dataset, R1 appeared more often in the free conversation and closing parts. This suggests that caregivers were actively listening to children's utterances by repeating backchannels such as "uh-huh." E and L tokens were less frequent than in CEJC-Child, particularly in the quiz part. This may be because the caregiver was leading the quiz. These results indicate that caregivers used different types of backchannels depending on the dialogue part.

5.3 Results of Utterance Intention Comparison

Figure 2 shows the distribution of utterance intention frequencies.

For children's utterance intentions (Fig. 2 (a)), in all corpora and dialogue parts, statements (utterances expressing facts, opinions, or desires, such as "Cats like milk and fish") and vocalizations (utterances consisting of sounds without clear communicative functions, such as "mm") appeared frequently. More specifically, in the introduction and free conversation parts, many utterances were responses to questions, while in the quiz part, there were many questions—typically guesses of correct

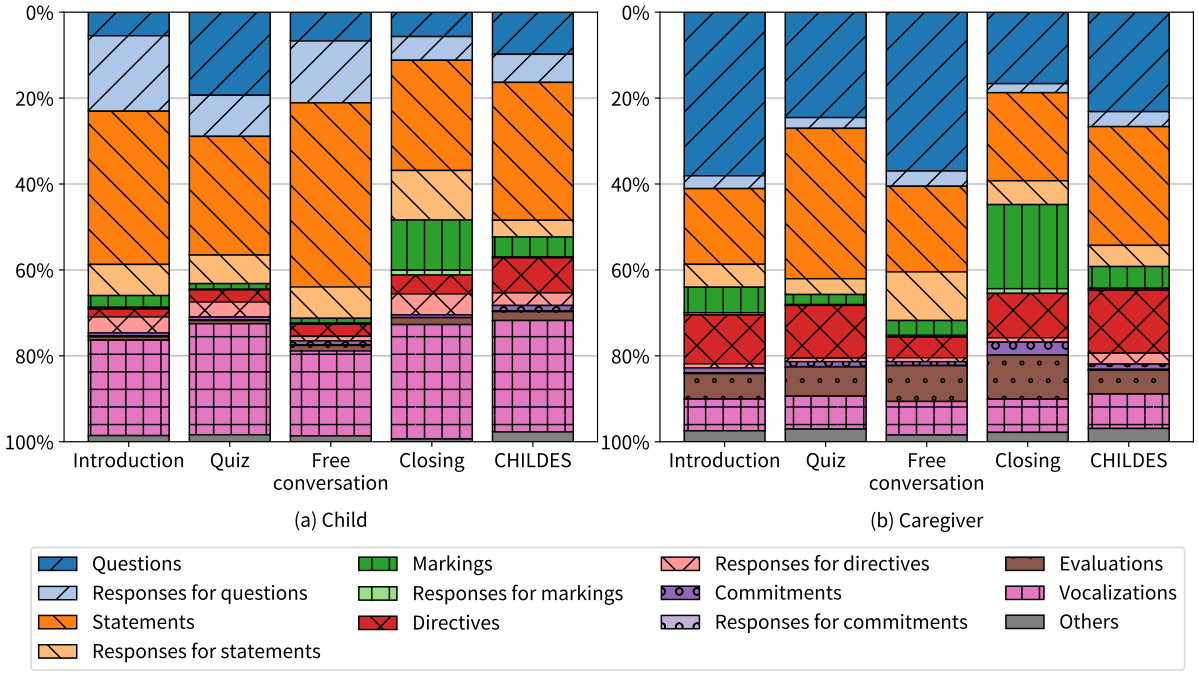

Figure 2: Frequency of utterance intentions.

answers (e.g., “A cat?”). In the closing part, markings (such as greetings like “Hello” or “Bye-bye,” and emotional reactions such as “Thank you” or “Wow”) and responses to statements appeared frequently.

In all corpora and dialogue parts, caregivers (Fig. 2 (b)) displayed a wider range of utterance intentions—such as directives, statements, questions, and evaluations—than children. This suggests that caregivers play a more controlling role in the dialogue, such as initiating actions, organizing content, and evaluating responses. Specifically, in the introduction and free conversation parts, questions were common; in the quiz part, statements were dominant; and in the closing part, markings and evaluations were frequent. In the RCDD, the frequency of evaluations increased toward the later parts of the dialogue. CHILDES included a relatively high frequency of statements and questions.

Figure 3 illustrates the transitions of utterance intentions in each part of the RCDD. The edge labels indicate transition probabilities.

From the overall structure, the introduction and free conversation parts appear relatively similar. These parts include frequent exchanges in which the caregiver’s question is followed by the child’s statement, to which the caregiver responds with an evaluation or comment. Such exchanges often correspond to basic adjacency pairs (Schegloff and Sacks, 1973; Schegloff, 2007). In contrast, the quiz part shows a different pattern. The core of the interaction is the caregiver’s statements and questions, which connect to the child’s statements or questions and then return again to the caregiver’s questions, forming a cycle of hint presentation, response, and subsequent hint presentation. This cyclical structure contrasts with the more balanced question–response–follow-up flow observed in the free conversation part, highlighting a clear asymmetry between instructional and conversational interaction modes.

The closing part shows distinct characteristics, with frequent transitions from the child’s markings to the caregiver’s markings. Overall, the caregiver’s questions and both child and caregiver statements occurred at roughly similar frequencies. Although the transition diagrams for CHILDES are omitted here for brevity, the transitions centered on caregiver statements and questions, and child statements, which connect bidirectionally with a variety of other utterance intentions. CHILDES appears to exhibit transitions similar to those observed in the quiz part.

5.4 Discussion

From the comparative results presented above, we obtained several insights for building a Japanese caregiver spoken dialogue system. First, the analysis of turn-taking revealed that child–caregiver

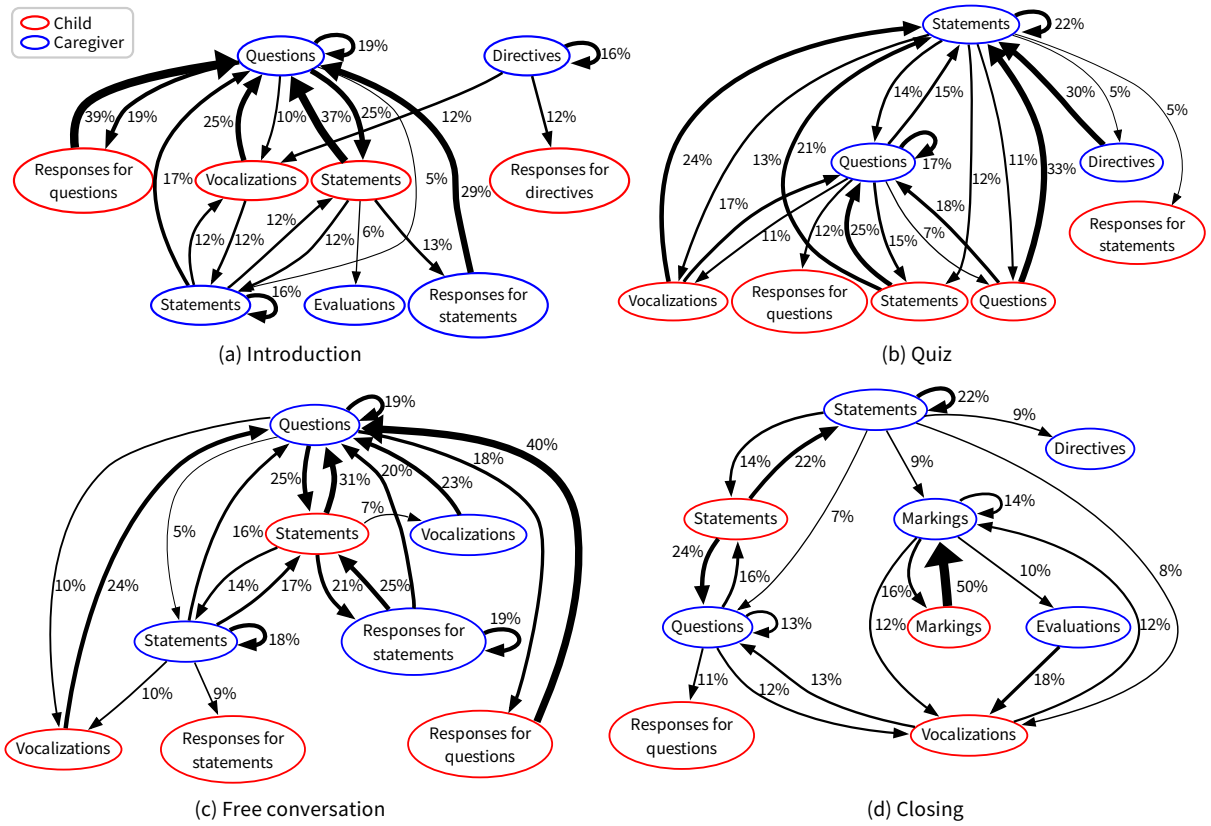

Figure 3: Transitions of utterance intentions by dialogue part.

dialogues exhibit little overlap between utterances. Therefore, while full-duplex systems (Défossez et al., 2024; Ohashi et al., 2025), which allow simultaneous speaking and listening, are desirable from the perspective of natural interaction, such a high level of bidirectionality may not be essential. A turn-based dialogue structure, which proceeds in a more controlled manner, may be appropriate for this context.

The analysis of backchannels showed that different dialogue parts require different types of backchannels. In particular, the quiz part requires expressive interjections to be frequently used.

The analysis of utterance intentions revealed that the distribution and transitions of intentions differ across dialogue parts. Current spoken dialogue models generally have limited controllability, and a single model may have difficulty handling such diverse interactions. Thus, preparing separate models for each dialogue part may be a practical approach at present.

In addition, in the free conversation part, many instances of the question–response–follow-up structure (Coulthard, 2014) were observed. Therefore, the system needs to continue the dialogue by evaluating and expanding on the child's responses.

6 Summary and Future Work

In this study, we collected a Japanese child–caregiver dialogue dataset (Remote Childcare Dialogue Data; RCDD) and analyzed it from multiple perspectives to identify distinctive features in it. On the basis of the findings, we also derived design guidelines for implementing spoken dialogue systems.

This study has several limitations. First, the attributes of the speakers are limited. Future work should include collecting corpora that cover a wider range of languages, ages, personality types, and interaction styles. Second, while this study focused on turn-taking, backchannels, and utterance intentions, we aim to examine other key factors in dialogue, such as prosodic information. Third, for LLM-based utterance intent annotation, we aim to evaluate the reliability of the approach using a larger amount of data. Finally, as a long-term goal, we plan to build a spoken dialogue system for remote childcare services based on the findings of this study. In developing such systems, careful consideration must be given to ethical issues, such as child safety and privacy protection.

7 Acknowledgments

This work was supported by JST Moonshot R&D Grant number JPMJMS2011. We would like to express our sincere gratitude to the users of NannyME for their generous cooperation in the collection of dialogue data.

References

Federico Biagi, Cristina Iani, and Luigi Biagiotti. 2025. The use of the social robot NAO in medical settings: How to facilitate interactions between healthcare professionals and patients with autism spectrum disorder. *Frontiers in Psychiatry*, 16:1675098.

Kübra Bodur, Mitja Nikolaus, Fatima Kassim, Laurent Prévot, and Abdellah Fourtassi. 2021. ChiCo: A multimodal corpus for the study of child conversation. In *Proceedings of the Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 158–163.

Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. 2023. Using video calls to study children's conversational development: The case of backchannel signaling. *Frontiers in Computer Science*, 5:1088752.

Marisa Casillas, Susan C. Bobb, and Eve V. Clark. 2016. Turn-taking, timing, and planning in early language acquisition. *Journal of Child Language*, 43(6):1310–1337.

Malcolm Coulthard. 2014. *An introduction to discourse analysis*. Routledge.

Mirjam de Haas, Paul Vogt, Rianne van den Berghe, Paul Leseman, Ora Oudgenoeg-Paz, Bram Willemsen, Jan de Wit, and Emiel Krahmer. 2022. Engagement in longitudinal child-robot language learning interactions: Disentangling robot and task engagement. *International Journal of Child-Computer Interaction*, 33:100501.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Yasuharu Den, Nao Yoshida, Katsuya Takanashi, and Hanae Koiso. 2011. Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *Proceedings of the 2011 International Conference on Speech Database and Assessments*, pages 168–173.

Alireza Esfandbod, Zeynab Rokhi, Ali F Meghdari, Alireza Taheri, Minoo Alemi, and Mahdieh Karimi. 2023. Utilizing an emotional robot capable of lip-syncing in robot-assisted speech therapy sessions for children with language disorders. *International journal of social robotics*, 15(2):165–183.

Tanya M Gallagher and Holly K Craig. 1982. An investigation of overlap in children's speech. *Journal of Psycholinguistic Research*, 11(1):63–75.

Dhia Elhak Goumri, Abhishek Agrawal, Mitja Nikolaus, Hong Duc Thang Vu, Kübra Bodur, Elias Emmar, Cassandre Armand, Chiara Mazzocconi, Shreejata Gupta, Laurent Prévot, Benoit Favre, Leonor Becerra-Bonache, and Abdellah Fourtassi. 2024. CHICA: A developmental corpus of child-caregiver's face-to-face vs. video call conversations in middle childhood. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 3153–3164.

Viktória Horváth and Valéria Krepsz. 2023. Temporal characteristics of child-adult conversations: Utterances and turn-taking. *Taikomoji kalbotyra*, (19):3–13.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G. Ward. 2016. Prediction and generation of backchannel form for attentive listening systems. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, pages 2890–2894.

Min Kyong Kim, Stefania Druga, Shaghayegh Esmaeili, Julia Woodward, Alex Shaw, Ayushi Jain, Jaida Langham, Kristy Hollingshead, Silvia B Lovato, Erin Beneteau, Jaime Ruiz, Lisa Anthony, and Alexis Hiniker. 2022. Examining voice assistants in the context of children's speech. *International Journal of Child-Computer Interaction*, 34:100540.

Hanae Koiso, Yuichi Ishimoto, Iseki Yuriko, Noriko Eguchi, Wakako Kashino, Yoshiko Kawabata, Mariko Tanaka, Yayoi Tanaka, and Ken'ya Nishikawa. 2025. Construction of the pilot version of the corpus of everyday Japanese conversation for child. In *Proceedings of the Thirty-first Annual Meeting of the Association for Natural Language Processing*, pages 3525–3528. (In Japanese).

Jacqueline M Kory-Westlund and Cynthia Breazeal. 2019. A long-term study of young children's rapport, social emulation, and language learning with a peer-like robot playmate in preschool. *Frontiers in Robotics and AI*, 6:81.

Jing Liu and Abdellah Fourtassi. 2024. Benchmarking LLMs for mimicking child-caregiver language in interaction. *arXiv preprint arXiv:2412.09318*.

Jing Liu, Mitja Nikolaus, Kübra Bodur, and Abdellah Fourtassi. 2022. Predicting backchannel signaling in child-caregiver multimodal conversations. In *Proceedings of the Companion publication of the 2022 international conference on multimodal interaction*, pages 196–200.

Brian Macwhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates.

Sayo Miyata. 2012. *Japanese CHILDES: The 2012 CHILDES manual for Japanese*.

http://www2.aasa.ac.jp/people/smiyata/CHILDESmanual/chapter01.html (in Japanese).

Jack Mostow, Greg Aist, Paul Burkhead, Albert Corbett, Andrew Cuneo, Susan Eitelman, Cathy Huang, Brian Junker, Mary Beth Sklar, and Brian Tobin. 2003. Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 29(1):61–117.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Mitja Nikolaus, Eliot Maes, Jeremy Auguste, Laurent Prévot, and Abdellah Fourtassi. 2022. Large-scale study of speech acts' development in early childhood. *Language Development Research*, 2(1):268–304.

Anat Ninio, Catherine E. Snow, Barbara A. Pan, and Pamela R. Rollins. 1994. Classifying communicative acts in children's interactions. *Journal of Communication Disorders*, 27(2):157–187.

Atsumoto Ohashi, Shinya Iizuka, Jingjing Jiang, and Ryuichiro Higashinaka. 2025. Towards a Japanese full-duplex spoken dialogue system. In *Proceedings of the 26th Interspeech Conference*, pages 1783–1787.

Nazerke Rakhymbayeva, Aida Amirova, and Anara Sandygulova. 2021. A long-term engagement with a social robot for autism therapy. *Frontiers in robotics and AI*, 8:669972.

Yoko Saikachi, Kazuki Watanabe, Takayuki Konishi, Naoko Ito, Ai Kanato, Yosuke Igarashi, Koki Miyazawa, Ken'ya Nishikawa, and Reiko Mazuka. 2013. Riken Japanese mother infant conversation corpus (R-JMICC) – compilation and recent findings of Japanese-specific prosodic and segmental characteristics in infant-directed speech–. In *Proceedings of the 3rd Workshop on Corpus-based Japanese Linguistics*, pages 383–392. (In Japanese).

Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge University Press.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

Wayne Ward, Ronald Cole, Daniel Bolaños, Cindy Buchenroth-Martin, Edward Svirsky, Sarel Van Vuuren, Timothy Weston, Jing Zheng, and Lee Becker. 2011. My science tutor: a conversational multimedia virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing*, 7(4).

Ying Xu, Joseph Aubele, Valery Vigil, Andres S. Bustamante, Young-Suk Kim, and Mark Warschauer. 2022. Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement. *Child Development*, 93:307–611.

8 Appendix

```
# Instructions
Your task is to assign a label to each utterance in a given dialogue. Follow the constraints and output format below.

# Constraints
For each utterance in the given dialogue, select the most appropriate label from the following 19 options, taking the context into account.

1. Directives
Utterances by which the speaker requests, proposes, or commands that the listener (or both listener and speaker) perform some action. This category also includes action-eliciting questions (e.g., "Will you do xxx?"), as well as dares, warnings, vocatives, and start signals.
Examples:
- Requests/proposals/presentation of action: Utterances requesting or proposing that the listener, or both listener and speaker, perform an action. "Let's clean up", "Let's go together"
- Yes/No questions about wishes or intentions (suggestive function): Yes/No questions that inquire about the listener's wishes or intentions and function as a suggestion. "Do you want to do it again?", "Do you want to sit?"
- Dares/challenges: Utterances that challenge the listener to perform an action. "Can you do it? Go ahead!"
- Warnings: Utterances that warn of danger. "It's hot", "Be careful!"
- Calls/vocatives: Utterances that attract the listener's attention by name, title, or exclamation. "Mom!", "Hey!"
- Start signals: Signals that initiate an action or coordinate timing. "Ready, set, go!", "One, two!"

2. Responses_for_Directives
Utterances that express agreement, refusal, or related responses to Directives.
Examples:
- Agreement: Agreeing to another's request or proposal and committing to carry it out. "Okay, I will", "Sure"
- Agreement to do for the last time: Agreeing that an action will be done for the last time. "Okay, this is the last one"
- Refusal: Expressing unwillingness to carry out another's request or proposal. "No", "I won't"
- Counter-suggestion: Offering an alternative as an indirect refusal. "Let's do it later", "Let's do this instead"
- Giving in: Accepting the other's insistence or refusal. "Alright, I'll stop"
- Response to a call: Answering a vocative and showing attentiveness. "What?", "Yes?"
- Giving reasons: Providing justification for a request, refusal, or prohibition. "It's dangerous, so no", "I'm tired, so I won't go"

3. Speech_Elicitations
Acts designed to elicit speech from the interlocutor, or responses to such acts. This category includes requests for imitation, completion, or vocal imitation of words or sentences.
Examples:
- Eliciting imitation: Eliciting imitation of a word or sentence by modeling or explicit command. "Say 'dog'"
- Eliciting completion of a sentence: Prompting the continuation of a sentence or word. "The moon is...?" (-> "pretty")
- Eliciting completion of a rote-learned text: Prompting completion of a memorized formula or routine. "Itadaki...?" (-> "masu")
- Eliciting onomatopoeic or animal sounds: Prompting the production of sound effects or animal noises. "Say 'woof woof'"

4. Responses_for_Speech_Elicitations
Acts produced in response to Speech_Elicitations.
Examples:
- Repetition/imitation: Repeating the other's utterance. Parent: "Banana" -> Child: "Banana"
- Completion response: Completing an utterance as requested. "xxx is...?", "Here!"
- Completion of rote text: Completing a formulaic expression. "Itadaki...", "masu!"

5. Commitments
Utterances expressing the speaker's own intentions, permissions, or prohibitions concerning future actions, that is, declarations about the speaker's own behavior.
Examples:
- Statement of intent: Expressing intention to carry out an action or describing an ongoing action. "I'll make this", "I'm cleaning up now"
- Request for permission: Asking permission to perform an action. "Can I eat this?", "May I touch it?"
- Promise: Committing oneself to a future action. "I'll go later"
- Threat: Warning that the speaker will carry out an undesirable action. "If you cry, I won't help you anymore"

6. Responses_for_Commitments
Utterances that grant permission, prohibit, or otherwise respond to Commitments.
Examples:
- Permission: Allowing the hearer to perform an action. "Okay, go ahead"
- Prohibition/protest: Forbidding or objecting to the hearer's action. "That's not allowed", "Don't touch it!"

7. Declarations
Acts by which a new social or factual state of affairs is created by the utterance itself.
Examples:
- Declaration: Creating a new state of affairs by declaration. "Today is a holiday", "That's the end!"
- Declaration of make-believe reality: Declaring an imagined reality in pretend play. "This is a castle", "You are the prince!"

8. Responses_for_Declarations
Utterances expressing agreement or disagreement with a Declaration.
Examples:
- Agreement with a declaration: Accepting another's declaration. "Yes, that's right"
- Disagreement with a declaration: Challenging the content of a declaration. "No, that's not a house"

9. Markings
Acts that socially mark the occurrence of events or express affective reactions. This category includes social routines such as thanking, greeting, apologizing, and celebrating.
Examples:
- Marking events: Expressing socially expected sentiments such as thanks, greetings, apologies, celebration, or marking the end of an action. "Thank you", "Sorry", "Yay!"
- Transfer of object: Marking the giving of an object to the hearer. "Here you are"
- Commiseration/empathy: Expressing sympathy for the hearer's misfortune or pain. "That hurt, didn't it?", "Poor thing"
- Expression of distress: Expressing pain or discomfort. "Ouch!", "No!"
```

```
- Expression of pleasure: Expressing positive emotion. "I'm happy!", "Yay!"
- Expression of surprise: Expressing surprise. "What!", "Wow!"
- Exhibiting attentiveness: Showing attention to the interlocutor. "Uh-huh, I'm listening"

10. Responses_for_Markings
Socially appropriate responses to Markings.

11. Statements
Utterances that state facts, opinions, or desires.
Examples:
- Declarative statements: Stating facts or information. "This is red", "It's raining"
- Wishes: Expressing desires. "I want to play soon", "I want to eat snacks"
- Counting: Producing number sequences. "One, two, three..."

12. Responses_for_Statements
Utterances expressing agreement or disagreement with Statements.
Examples:
- Agreement: Agreeing with the proposition of the prior utterance. "That's right", "Uh-huh"
- Disagreement: Disagreeing with the prior utterance. "No", "That's not it"

13. Questions
Utterances that request information, including wh-questions, Yes/No questions, alternative questions, and confirmation questions.
Examples:
- wh-questions: Questions seeking information in wh-form. "What is this?", "Where are we going?"
- Yes/No questions: Questions requesting affirmation or negation. "Do you like it?", "Is it done?"
- Limited-alternative questions: Questions presenting alternatives. "Red or blue?"
- Eliciting questions: Questions prompting brief responses. "Hm?", "Eh?"
- Aggravated questions: Repetition of a question with a negative stance. "Did you make a mess again?"

14. Responses_for_Questions
Utterances produced in response to Questions.
Examples:
- Answers to wh-questions (sentential): Answering a wh-question with a sentence. "This is an apple"
- Affirmative answers: Affirmative responses to Yes/No questions. "Yes", "Yeah"
- Negative answers: Negative responses to Yes/No questions. "No", "Nope"
- Answering with a wh-question: Responding with another wh-question. "What is this?" -> "Which one?"
- Answering with a Yes/No question: Responding with a Yes/No question. "Do you like it?" -> "Do you, Mom?"
- Answers to limited-alternative questions: Selecting one of the alternatives. "Blue!"
- Intentionally non-satisfying answers: Providing an incomplete response. "Um...", "I don't know"
- Refusal to answer: Expressing unwillingness to answer. "I won't say", "It's a secret"

15. Performances
Utterances produced as part of rule-governed games or activities, including in-game verbal moves and recitation.
Examples:
- Verbal moves in activities: Utterances produced according to the rules of a game or activity. "Pass!", "Goal!"
- Reading/recitation: Reading aloud written text. "Once upon a time..."

16. Evaluations
Utterances expressing positive or negative evaluations of the hearer's or speaker's actions, including praise, criticism, and
reprimands.
Examples:
- Praise for actions: Praising nonverbal behavior. "You did it well!"
- Exclamations of enthusiasm or surprise: Praising with excitement or delight. "Wow, great!", "You did it!"
- Pointing out errors: Indicating mistakes in action. "That's wrong there", "Once more"
- Approval of appropriate behavior: Positively evaluating correct or desirable behavior. "Good", "Do it like that"
- Negative evaluation/scolding: Expressing disapproval of inappropriate behavior. "That's not okay", "Don't do that"
- Expression of displeasure: Exclaiming dissatisfaction or aversion. "I don't like it!", "Enough!"

17. Demands_for_Clarification
Acts requesting repetition or clarification of a prior utterance.
Examples:
- Requests for repetition: Requesting that the interlocutor repeat an utterance. "Huh? What did you say?"

18. Text_Editing
Acts that correct another's erroneous utterance by providing the appropriate linguistic form.
Examples:
- Correction: Replacing an incorrect linguistic form with the correct one. Child: "jo-ju" -> Parent: "jo-zu, right"

19. Vocalizations
Word-like or nonword vocalizations with no clear communicative function, including unintelligible sounds.
Examples:
- Word-like/unintelligible vocalizations: Vocalizations without identifiable function or meaning. "Ah", "mma"

# Output Format
{"utterance_id":0,"label":"Markings"}
{"utterance_id":1,"label":"Responses_for_Markings"}
{"utterance_id":2,"label":"Questions"}
...

# Task
## Given Dialogue
$dialogue
```

Figure 4: Prompt used for estimating utterance intentions. The text was originally in Japanese and was translated by the authors.