On the Structure of Address in Multi-Party Dialogue: From Discrete Labels to Continuous Levels

Taiga Mori, Koji Inoue, Divesh Lala, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Japan
Correspondence: mori@sap.ist.i.kyoto-u.ac.jp

Abstract

In multi-party dialogues between a dialogue system and multiple users, identifying to whom an utterance is addressed is a key challenge. Prior work has typically treated addressee detection as a multi-class classification task, selecting a single label representing an individual participant or the group. This formulation assumes that address is inherently discrete and has primarily been used for predicting turn-taking. In this paper, we revisit this assumption by analyzing address as a continuous phenomenon. Using a multi-party human dialogue corpus annotated by multiple annotators, we construct both binary address labels derived from majority-vote addressee labels and continuous address levels inferred from annotator judgments using a latent-variable model. We then examine how these representations relate to turn-taking as well as listener behaviors, including gaze and backchannels. Our results show that, in addition to turn-taking, both gaze and backchannels are associated with address. Furthermore, models using continuous address levels achieve better predictive fit than those using discrete labels, suggesting that address may exhibit graded structure. Finally, we discuss the future directions of addressee detection research based on the findings of this study.

1 Introduction

Identifying to whom an utterance is addressed is one of the fundamental challenges in multi-party dialogues involving multiple users and a dialogue system. As discussed below, most prior work on addressee detection has formulated addressee inference as a classification problem. In human-human-computer interaction, the task is typically treated as a binary classification problem that determines whether a user's utterance is addressed to the system or to another human participant. In multi-party human dialogue, by contrast, it is commonly formulated as selecting a single addressee from among labels corresponding to individual participants or the group as a whole. Moreover, in dialogue-system research, addressee detection has primarily been used for predicting turn-taking.

Such conventional formulations rely on the simplified assumption that each utterance is directed toward a single discrete addressee or toward a predefined category such as the whole group. However, there is little reason to assume that naturally occurring dialogue is always so simple, and address may in fact be more complex. Under such formulations, disagreement among annotators is typically treated as mere error. In addition, address is often modeled as an isolated phenomenon, and relatively little work has examined its relationship to listener behaviors that are likely to be associated with it, or attempted to model these phenomena jointly.

From this perspective, the present study analyzes addressee annotations assigned by multiple annotators to multi-party human dialogues and shows that address may not be a purely discrete phenomenon, but instead may exhibit graded structure. We also introduce *address level* as a new concept for addressee modeling, representing addressivity as a continuous property. Furthermore, we show that address is related not only to turn-taking but also to other listener behaviors.

The contributions of this study are threefold. First, we provide empirical grounds for reformulating addressee modeling not as a discrete classification problem but as a continuous estimation problem. Second, we propose an operationalization that infers participant-wise continuous representations from categorical annotations. Third, we show that these continuous addressivity measures correlate with observable listener behaviors, thereby supporting their validity as a representation of interactional structure.

This paper is structured as follows. Section 2 organizes previous research on addressee detection and situates the present study within that literature. Section 3 describes the dataset and annota-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 668–679
August 2–5, 2026.

tion scheme and presents a preliminary analysis of the ambiguity inherent in addressee annotation. Section 4 explains how address labels and address levels were derived, as well as how turn-taking, listener gaze, and backchannel behavior were modeled. Section 5 presents the quantitative results and a qualitative analysis. Section 6 concludes the paper with a discussion of implications and future directions. Limitations are discussed in a separate section.

2 Related Work

Most prior work on addressee detection has formulated the task as a discrete classification problem. Early studies on meeting dialogues treated addressee detection as assigning discrete labels to dialogue acts, distinguishing utterances addressed to an individual participant from those addressed to the group; when an individual was addressed, the label specified the participant ID (Jovanovic et al., 2006; op den Akker and op den Akker, 2009). In text-based multi-party dialogue research, the task has often been formulated as selecting one addressee from the candidate interlocutors appearing in the dialogue context (Ouchi and Tsuboi, 2016; Zhang et al., 2018; Sato et al., 2018; Gu et al., 2021; Zhu et al., 2023). In human-human-computer settings, by contrast, the label space is often binary, distinguishing system-addressed speech from speech addressed to another human participant (Baba et al., 2011; Shriberg et al., 2012; Tsai et al., 2015; Nakano et al., 2013).

Regarding the timing of prediction, many studies perform inference after the target turn has been fully observed. Early work treated the task as assigning addressee labels to dialogue acts or manually segmented utterances (Jovanovic et al., 2006), and this utterance-level formulation remains common in recent studies. For example, Ouchi and Tsuboi (2016) jointly predict the addressee and response given the full context and target utterance, while Penzo et al. (2024) evaluate addressee recognition for the final turn in multi-party dialogues. Thus, a common paradigm is post-hoc prediction at the utterance level, rather than incremental estimation during the progression of an utterance.

Furthermore, addressee detection has often been treated as an auxiliary task for turn-taking or response selection. In human-human-computer interaction, it is used to determine whether the system should respond to an utterance (Shriberg et al., 2012; Tsai et al., 2015). In text-based multi-party dialogue, addressee selection is commonly integrated with response selection (Ouchi and Tsuboi, 2016; Zhang et al., 2018), and MPC-BERT (Gu et al., 2021) also treats addressee recognition as one component among multiple dialogue understanding tasks. In this way, addressee detection has often been used as an intermediate representation for turn-taking and response generation, rather than being studied as an independent dialogue behavior.

From a methodological perspective, the field has undergone a progressive evolution. Early studies primarily relied on statistical methods such as SVMs and logistic regression using features derived from gaze, prosody, and lexical cues (Jovanovic et al., 2006; Shriberg et al., 2012; Tsai et al., 2015). Subsequently, neural network-based approaches were introduced to model dialogue context and speaker relationships (Ravuri and Stolcke, 2014; Ouchi and Tsuboi, 2016; Zhang et al., 2018). Later, multimodal deep learning approaches incorporating visual and contextual information were proposed (Akhtiamov et al., 2017; Le et al., 2018). More recently, recent work has also explored pretrained language models (Gu et al., 2021) and inference-based approaches using LLMs and MLLMs (Zhu et al., 2023; Penzo et al., 2024; Inoue et al., 2025; Mori et al., 2026).

Along with these methodological advances, predictive performance has steadily improved. While early studies reported accuracies of around 80% (Jovanovic et al., 2006), recent approaches using MLLMs have achieved substantially higher performance. For instance, Mori et al. (2026) reports a Macro-F1 score exceeding 0.9 for cases where the single next speaker is identifiable. Moreover, as performance improves, recent research has begun to address additional challenges such as multilingual settings (Sato et al., 2018) and robustness under noisy conditions (Zhu et al., 2023).

In this study, we revisit these underlying assumptions by analyzing addressee annotations from multiple annotators, as well as the relationships between addressing behavior, turn-taking, listener gaze, and backchannels. Based on these analyses, we discuss future directions for addressee detection research.

Figure 1: A snapshot from the Teidan corpus.

3 Dataset

3.1 Corpus Description

In this study, we used the dataset introduced in prior work by Inoue et al. (2025). This corpus contains casual discussions in Japanese among triads of members of the same laboratory. Twelve groups each conversed on three topics, resulting in a total of 36 recorded dialogues. Figure 1 shows a screenshot from one of the recorded dialogues. The average duration of a dialogue is 6 minutes and 4 seconds, and the total duration is 3 hours, 38 minutes, and 27 seconds. The data are annotated manually with turn and backchannel information. A turn is defined as a segment during which one participant speaks continuously. By contrast, short reactive utterances that mainly signal listening or prompt another participant to continue are annotated as backchannels rather than treated as independent turns. These include brief responses such as *hai* ("yes"), *ee* ("yes"), *un* ("yeah/uh-huh"), and *fuun* ("I see"), short repetitions of a prior utterance, and standalone interjections such as *aa* ("ah") or *ee* ("oh") when they are not followed by a more substantive utterance. The exception is that "hai" is treated as a turn when it functions as an answer to a question. The average number of turns per dialogue is 86.7, and the total number of turns is 3,121. Hereafter, for convenience, the three participants are referred to as A, B, and C.

3.2 Addressee Annotation

Based on Kadota et al. (2024), each turn was assigned one of eight addressee labels. Table 1 summarizes the label set used in this study.

Label	Meaning
A, *B*, *C*	addressed to one specific participant
E	everyone
W	whoever
N	none
O	other
U	unknown

Table 1: Addressee labels used in the annotation.

First, the authors conducted a pilot annotation on one dialogue. Based on this pilot, three annotators annotated the 3,078 turns in the remaining 35 dialogues. For each dialogue, three annotators performed the annotation independently. They were instructed to use multimodal information when judging the addressee, specifically explicit address terms, person references, register choice, references to another participant's prior experience or knowledge, links to prior talk such as disagreement markers (e.g., *but*), deictic expressions and repetitions referring to previous utterances, whether the utterance was a response to a prior question, gaze direction, and pointing or other gestures. They were also instructed not to take into account any information occurring after the target turn.

3.3 Gaze Annotation

The gaze annotation classified each segment of a dialogue into one of three categories: the two participants other than the target participant (i.e., two of A, B, and C), or elsewhere (O). As with the addressee annotation, one dialogue was first annotated on a trial basis by students in our laboratory, and the remaining 35 dialogues were then annotated. Due to technical problems, some cameras stopped before the end of the dialogue in six dialogues; for those intervals after a camera stopped, no gaze annotation is available for the corresponding participant. In addition, intervals during which the gaze was shifting from one participant to another were classified as O, and participant labels were assigned only to intervals in which the annotator could judge that the gaze was clearly directed at that person.

3.4 Preliminary Analysis

To investigate addressee ambiguity, we analyzed the annotations for the 35 dialogues produced by the crowd workers. Table 2 presents the frequency of each label. Among the participant-specific labels *A*, *B*, and *C*, *B* occurred somewhat less often

Raw label	Count
A	1932
B	1390
C	1751
E	3529
N	247
O	3
U	48
W	334

Table 2: Counts of raw addressee labels

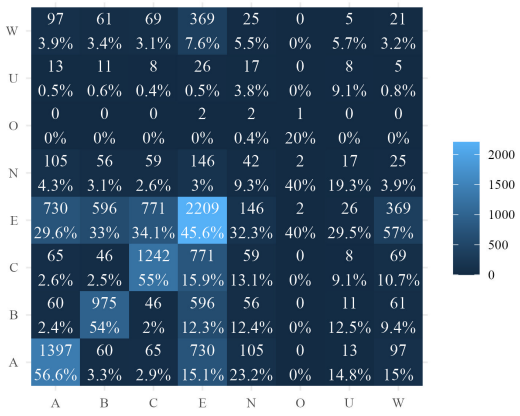

Figure 2: Co-occurrence heatmap of annotator address labels.

than the other two, but all three appeared between 1,000 and 2,000 times. The most frequent label was *E*, which occurred more than 3,500 times. By contrast, *W*, *N*, and *O* were relatively infrequent. These results suggest that most utterances were judged to be clearly addressed either to a specific individual or to all participants. This is likely because the data consist of discussion-style dialogues, in which many utterances either question or challenge another participant's opinion or present the speaker's opinion to the group as a whole. As a measure of inter-annotator agreement, we calculated Fleiss' kappa. For the 3,078 turns annotated by three annotators, the agreement was $\kappa = 0.52$, indicating moderate agreement. This suggests that the addressee annotation was reasonably consistent overall, while still leaving room for ambiguity.

Next, to examine which labels were prone to ambiguity, Figure 2 shows a co-occurrence heatmap of label pairs. Co-occurrence counts were computed as follows: for example, if the three annotators labeled a turn as *A*, *A*, and *B*, the pair *AA* was counted once and the pair *AB* was counted twice. The color intensity and the upper number in each cell indicate the raw count, while the lower number shows the proportion of that pair among all pairs involving the column label; these proportions sum to 100% within each column.

For *A*, *B*, and *C*, co-occurrence with the same label was the most frequent, exceeding 50% in each case. The second most frequent co-occurrence was with *E*, at around 30%. A complementary pattern was observed for *E*: co-occurrence with itself was the most frequent, at about 45%, but co-occurrence with *A*, *B*, and *C* was also substantial, at around 15% each. These results indicate that judgments distinguishing utterances addressed to a specific individual from those addressed to the group were fairly stable, but not perfectly so. For *N*, the most frequent co-occurrence was with *E*, at 32%. One possible reason is that both *E* and *N* share the property that the two other participants are treated similarly with respect to address. However, co-occurrence of *N* with itself was only about 9%, whereas co-occurrence with *E* was about 3.5 times larger; given that *E* itself occurred more than 14 times as often as *N*, this does not appear to be excessive relative to chance. Like *N*, *W* most frequently co-occurred with *E*, with a frequency about 17 times higher than its co-occurrence with itself. Since the difference in their marginal frequencies is only about tenfold, the co-occurrence between *W* and *E* is strikingly high even after taking label imbalance into account. This tendency was also noted in the prior study on which our label set was based (Kadota et al., 2024). Unlike *N*, *W* still implies that someone is being addressed, although not a specific individual, and in that respect it resembles *E*. No meaningful pattern was found for *O* or *U* because they were too rare.

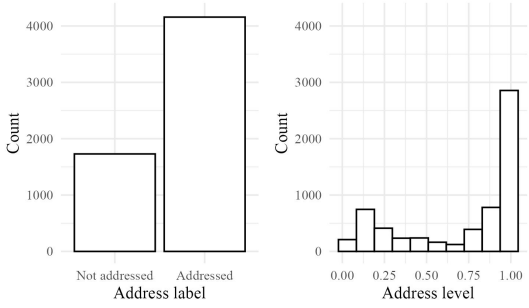

Figure 3: Distribution of address labels and address level.

4 Method

4.1 Address Label and Address Level

Section 3.4 showed that addressee annotation contains a certain degree of ambiguity. In this section, we argue that this ambiguity does not simply reflect annotator-specific errors, but may instead reflect ambiguity inherent in the act of addressing itself. We also examine whether address is related not only to turn-taking but also to other listener behaviors. To this end, we construct two representations of addressivity: an *address label*, which follows the conventional discrete distinction between addressed and not addressed, and an *address level*, which captures addressivity as a continuous property. We then compare these two representations.

We first describe how the address label was constructed. Based on the results in Section 3.4, *W* was merged into *E*. Next, the label *U*, which indicates that the addressee could not be determined, was excluded. For the remaining six labels, a majority vote was taken for each turn, and when a majority label existed, it was used as the aggregated addressee label for that turn. Turns for which all annotators assigned *U*, or for which no majority label was obtained, were excluded from this analysis. In our data, the former case did not occur, whereas 134 turns fell into the latter category and were excluded. Then, for each of the two listeners in each turn, an *addressed* label was assigned if the majority addressee matched that listener (i.e., *A*, *B*, or *C*) or included that listener (i.e., *E*). Conversely, a *not addressed* label was assigned if the majority addressee matched the other listener, or if the utterance was labeled as monologic speech (*N*) or as being addressed to an entity outside the participants (*O*). This binary value was used as the address label for each listener in each turn.

Next, we describe how the address level was estimated. Unlike the address label, the address level was treated as a latent continuous variable inferred from the labels assigned by the annotators. For each turn i, we assumed two latent addressivity values, θ_{i1} and θ_{i2}, corresponding to the two listeners of the current speaker. These values represent the degree to which the utterance is addressed to listener 1 and listener 2, respectively. They were constrained to the unit interval by applying a logistic transformation to normally distributed latent scores:

$$\theta_{i1} = \text{logit}^{-1}(\eta_{i1}), \quad (1)$$

$$\theta_{i2} = \text{logit}^{-1}(\eta_{i2}). \quad (2)$$

The annotator labels were modeled as categorical observations generated from the latent addressivity vector $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})$. As with the address label, *W* was merged into *E*, labels referring to the current speaker and *O* were merged into *N*, and *U* was excluded from the likelihood. The remaining labels were associated with ideal points in the two-dimensional addressivity space: *L1* with $(1, 0)$, *L2* with $(0, 1)$, *E* with $(1, 1)$, and *N* with $(0, 0)$. Here, *L1* and *L2* denote the two listeners of the current speaker. We treat addressivity as listener-wise directedness rather than as a finite resource to be divided among listeners; therefore, *E* was represented as $(1, 1)$ rather than as a normalized allocation such as $(0.5, 0.5)$. For annotator j and label c, the probability of the observed label y_{ij} was defined as

$$\Pr(y_{ij} = c) = \text{softmax}_c \left[\alpha_{jc} - \lambda_c \left\| \boldsymbol{\theta}_i - \mathbf{v}_c \right\|^2 \right], \quad (3)$$

where $\mathbf{v}_c$ is the ideal point for label c, α_{jc} is an annotator-specific label bias, and λ_c is a label-specific discrimination parameter. Thus, labels whose ideal points were closer to the latent addressivity vector were assigned higher probabilities, while allowing annotators to differ in their tendencies to use particular labels. Although annotators may also differ in label-specific discrimination, the discrimination parameters were shared across annotators to avoid overparameterization, given that each turn was annotated by only three annotators. Annotator-level differences in label-use tendencies were instead captured by the bias parameters α_{jc}.

The latent-address model was estimated by Bayesian sampling using R version 4.6.0 and `cmdstanr` version 0.9.0. We used four chains, with 2,000 sampling iterations per chain and 1,000 warm-up iterations. For each turn-listener pair, the posterior median of the corresponding latent addressivity value was used as the address level in the subsequent analyses.

Figure 3 shows the distributions of the address labels and address levels. For the address labels, the number of *addressed* instances is approximately twice that of *not addressed*. This is consistent with the original label distribution, in which *E* was the most frequent label, followed by the participant-specific labels (*A*, *B*, and *C*). For the address levels,

many observations are concentrated near the high ends of the scale, but intermediate values are also observed, indicating the existence of partially addressed states.

4.2 Modeling

Next, we analyzed the relationships between these two address measures and turn-taking, listener gaze, and backchannel behavior. Each sample represented the behavior of each listener in each turn, yielding hierarchical data nested within turns, dialogues, speakers, and listeners. Accordingly, we modeled the data using generalized linear mixed-effects models and estimated the parameters by Bayesian sampling. The mixed-effects models were fitted using the `brms` package version 2.23.0. For all models, we used four chains, with 2,000 sampling iterations per chain and 1,000 warm-up iterations. As weakly informative priors, parameters defined over the entire real line, such as intercepts and regression coefficients, were assigned normal distributions with mean 0 and standard deviation 10, whereas non-negative parameters, such as sd and sigma, were assigned half-Cauchy distributions with location 0 and scale 10. In all models, the fixed effect was either address label or address level, and the random effects were random intercepts for turn, dialogue session, speaker, and listener. These random intercepts were defined as follows:

$$
\begin{aligned}
u_{turn,i} &\sim \mathcal{N}(0, \sigma_{\text{turn}}) \\
u_{session,i} &\sim \mathcal{N}(0, \sigma_{\text{session}}) \\
u_{speaker,i} &\sim \mathcal{N}(0, \sigma_{\text{speaker}}) \\
u_{listener,i} &\sim \mathcal{N}(0, \sigma_{\text{listener}})
\end{aligned}
$$

Convergence was assessed using diagnostics such as $\hat{R}$, Bulk ESS, and trace plots, and model comparisons were conducted using the Expected Log Predictive Density (ELPD) estimated with the `loo` function. ELPD evaluates out-of-sample predictive performance, with larger values indicating better expected predictive fit. To make the ELPD comparisons between the address-label and address-level models valid, the two models for each outcome were fit to the same set of observations. Therefore, turns for which no majority-vote address label was obtained were also excluded from the address-level models.

Turn-taking was modeled as follows.

$$y_i \sim \text{Bernoulli}(p_i)$$

$$
\begin{aligned}
\text{logit}(p_i) = \beta_0 &+ \beta_1 \, x_{\text{address\_label},i} + u_{turn,i} \\
&+ u_{session,i} + u_{speaker,i} + u_{listener,i}
\end{aligned}
\tag{4a}
$$

$$
\begin{aligned}
\text{logit}(p_i) = \beta_0 &+ \beta_1 \, x_{\text{address\_level},i} + u_{turn,i} \\
&+ u_{session,i} + u_{speaker,i} + u_{listener,i}
\end{aligned}
\tag{4b}
$$

Here, y_i is a binary variable indicating whether the focal listener in the turn corresponding to the i-th observation became the next speaker, and it is assumed to follow a Bernoulli distribution with parameter p_i. The predictors $x_{\text{address\_label},i}$ and $x_{\text{address\_level},i}$ denote the address label and address level for the focal listener, respectively. The random intercepts $u_{turn,i}$, $u_{session,i}$, $u_{speaker,i}$, and $u_{listener,i}$ represent variation associated with the turn, dialogue session, current speaker, and focal listener. The total number of observations was 5,820, calculated as the number of turns multiplied by the two listeners, excluding turns near the end of a dialogue for which no next speaker existed.

Listener gaze was modeled as follows.

$$y_i \sim \text{Beta}(\mu_i, \phi)$$

$$
\begin{aligned}
\text{logit}(\mu_i) = \beta_0 &+ \beta_1 \, x_{\text{address\_label},i} + u_{turn,i} \\
&+ u_{session,i} + u_{speaker,i} + u_{listener,i}
\end{aligned}
\tag{5a}
$$

$$
\begin{aligned}
\text{logit}(\mu_i) = \beta_0 &+ \beta_1 \, x_{\text{address\_level},i} + u_{turn,i} \\
&+ u_{session,i} + u_{speaker,i} + u_{listener,i}
\end{aligned}
\tag{5b}
$$

Here, y_i is the proportion of time during the turn corresponding to the i-th observation for which the focal listener was looking at the current speaker, and it is assumed to follow a Beta distribution with mean parameter μ_i. This distribution was used because the outcome is a proportion bounded between 0 and 1. Because the data include values of 0 and 1, the observed proportions were adjusted using the method of Smithson and Verkuilen (2006) before fitting the Beta model. Specifically, this transformation shrinks the observed proportions slightly toward the interior of the unit interval, so that values of 0 and 1 are mapped to values within $(0, 1)$ while preserving their relative ordering. The total number of observations was 5,744, excluding intervals for which gaze annotation was unavailable due to technical problems.

Backchannel was modeled as follows.

$$y_i \sim \text{Poisson}(\mu_i)$$

$$\begin{aligned}\log(\mu_i) = \beta_0 + \beta_1\, x_{\text{address\_label},i} \\ + \log(\text{duration}_i) + u_{turn,i} \\ + u_{session,i} + u_{speaker,i} + u_{listener,i}\end{aligned} \tag{6a}$$

$$\begin{aligned}\log(\mu_i) = \beta_0 + \beta_1\, x_{\text{address\_level},i} \\ + \log(\text{duration}_i) + u_{turn,i} \\ + u_{session,i} + u_{speaker,i} + u_{listener,i}\end{aligned} \tag{6b}$$

Here, y_i is the number of backchannels produced by the focal listener during the turn corresponding to the i-th observation, and it is assumed to follow a Poisson distribution. This distribution was used because the outcome is a count variable. The expected count is modeled as a function of fixed and random effects, with the duration of the turn included as an offset term. Accordingly, the model effectively estimates the rate of backchannel production rather than the raw count alone. Very short turns were excluded from the analysis because they often consisted of incomplete utterances or similarly brief segments for which backchanneling is unlikely in principle. Consistent with this, more than 90% of turns shorter than one second had zero backchannels. After excluding these short turns, the total number of observations used for the backchannel analysis was 4,920.

5 Results

5.1 Next Speaker

Figure 4 shows the posterior conditional effects of addressivity in Models 1a and 1b. Full posterior summaries of the fixed and random effects are reported in Appendix A. As shown in the figure, both models indicate that listeners with higher addressivity are more likely to become the next speaker. In Model 1a, the coefficient for *addressed* was positive (Est. = 0.72, 95% CI [0.59, 0.84]), and in Model 1b, the coefficient for *addressLevel* was also positive (Est. = 1.06, 95% CI [0.89, 1.23]), indicating clear positive effects of addressivity on next-speaker selection. The 95% credible intervals for both coefficients excluded zero. Table 3 reports the leave-one-out comparison. Model 1b, which uses address level, provides better predictive fit than Model 1a. This suggests that next-speaker selection is better captured by a continuous representation of addressivity than by a discrete address label alone. In other words, listeners who would be categorized as not addressed under a discrete formulation may still become the next speaker, whereas listeners categorized as addressed do not always do so.

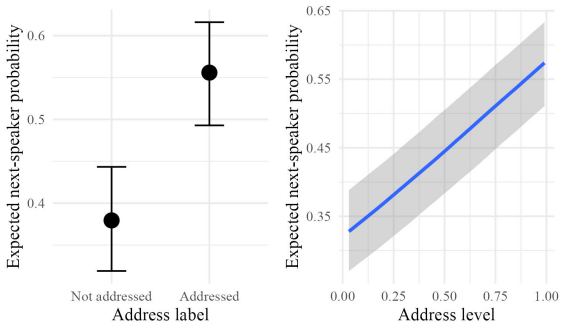

Figure 4: Posterior conditional effects of addressivity on next-speaker selection.

Model	elpd_diff	se_diff
(1b)	0.0	0.0
(1a)*	-14.4	4.9

Table 3: Leave-one-out comparison for the next-speaker models. Values are shown relative to the best model. An asterisk (*) indicates a substantial difference in predictive fit, defined here as $|\text{elpd\_diff}| > 1.96 \times \text{se\_diff}$.

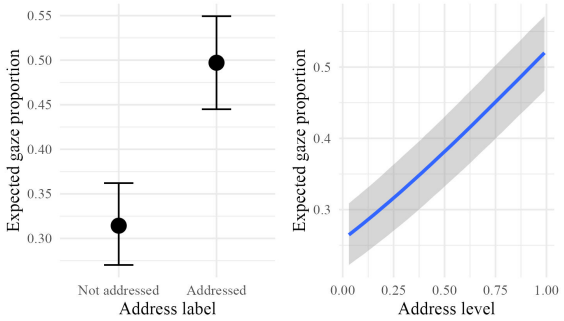

Figure 5: Posterior conditional effects of addressivity on listener gaze toward the current speaker.

5.2 Gaze

Figure 5 shows the posterior conditional effects of addressivity in Models 2a and 2b. Full posterior summaries of the fixed and random effects are reported in Appendix B. As shown in the figure, both models indicate that listeners with higher addressivity show higher gaze proportions toward the current speaker. In Model 2a, the coefficient for *addressed* was positive (Est. = 0.77, 95% CI

Model	elpd_diff	se_diff
(2b)	0.0	0.0
(2a)*	-39.2	7.9

Table 4: Leave-one-out comparison for the listener-gaze models. Values are shown relative to the best model.

Model	elpd_diff	se_diff
(3b)	0.0	0.0
(3a)*	-3.5	1.6

Table 5: Leave-one-out comparison for the backchannel models. Values are shown relative to the best model.

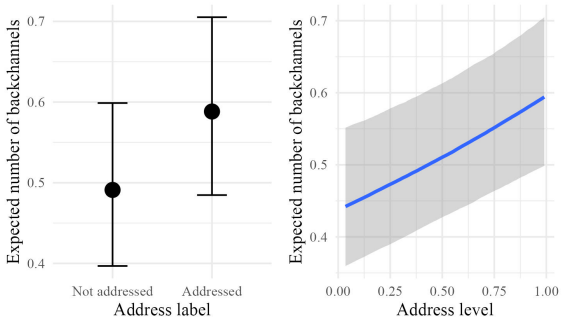

Figure 6: Posterior conditional effects of addressivity on listener backchannel production.

[0.69, 0.85]), and in Model 2b, the coefficient for *addressLevel* was also positive (Est. = 1.15, 95% CI [1.04, 1.26]), indicating clear positive effects of addressivity on listener gaze. Table 4 reports the leave-one-out comparison. Model 2b, which uses address level, provides better predictive fit than Model 2a. This suggests that listeners with higher addressivity tend to look at the speaker for longer periods of time, and that listener gaze is better captured by a continuous representation of addressivity than by a discrete address label alone. In other words, gaze toward the current speaker varies in a graded manner with the strength of addressivity, rather than being fully explained by a binary addressed/not-addressed distinction.

5.3 Backchannel

Figure 6 shows the posterior conditional effects of addressivity in Models 3a and 3b. Full posterior summaries of the fixed and random effects are reported in Appendix C. As shown in the figure, both models indicate that listeners with higher addressivity tend to produce more backchannels. In Model 3a, the coefficient for *addressed* was positive (Est. = 0.18, 95% CI [0.08, 0.28]), whereas in Model 3b, the coefficient for *addressLevel* was also positive (Est. = 0.31, 95% CI [0.17, 0.45]), indicating clear effects of addressivity on backchannel production. The 95% credible intervals for both coefficients excluded zero. Table 5 reports the leave-one-out comparison. Model 3b, which uses address level, provides better predictive fit than Model 3a. This suggests that listeners with higher addressivity tend to produce more backchannels, and that backchannel production is better captured by a continuous representation of addressivity than by a discrete address label alone. However, at the same time, the improvement is smaller than for next-speaker selection and listener gaze, suggesting that backchanneling is influenced not only by addressivity but also by additional factors.

5.4 Qualitative Analysis

In the following excerpt, the participants are discussing what means of transportation they would use if their club were to travel to Tokyo. Due to space limitations, we present only an English translation that is as faithful as possible to the original Japanese.

In line 01, C asks what they are supposed to do in Tokyo. While asking this question, C gazes at B, thereby strongly addressing B. Indeed, the next speaker in line 02 is B. Since line 01 is a question, that is, the first pair part of an adjacency pair, an answer as the second pair part is conditionally relevant in the next turn. However, what B produces in line 02 is another question. This question is not unrelated to C's question in line 01. While gazing at A, B asks whether it had been decided in the first place what they would do in Tokyo. In other words, B questions the presupposition of C's question. A then answers "shopping" in line 03. Interestingly, although A's answer is sequentially occasioned by B's question, A gazes at C and, in terms of content, answers C's question. For this reason, two annotators assigned the label C to this utterance. However, when observing the interaction, we do not get the impression that B's question is being ignored. This is because B's question concerns the presupposition of C's question, and answering C's question also provides an affirmative answer to B's question. In other words, A's utterance functions as an answer, or second pair part, not only to C's question but also to B's question. Thus, although the primary address is directed toward C, there is also a weaker degree of address toward B. Presumably

for this reason, the remaining annotator assigned the label E.

Excerpt1

01 C: Wait, what were we supposed to do in Tokyo again?

02 B: Was there something we were supposed to do in Tokyo?

03 A: Shopping.

6 Conclusions

The findings of this study have several implications for future research on addressee detection. First, the conventional framework of aggregating annotations into a single label and formulating the task as multiclass classification may oversimplify the act of addressing and discard important aspects of the phenomenon. In contrast, preserving variation across annotators and representing address continuously may provide a more appropriate way of modeling the phenomenon. The advantages of modeling address continuously are not limited to descriptive adequacy, but are also practically relevant for dialogue systems. For example, in conventional multiclass classification, utterances addressed to the whole group are collapsed into a single label, which by itself may not provide sufficient information for deciding turn allocation. By contrast, a continuous representation of address can capture subtle differences across participants even within cases that would previously have been treated uniformly as group-addressed. Moreover, a framework that estimates participant-wise address levels is less dependent on the number of participants in a dialogue and may therefore support more scalable model designs.

Second, the finding that address is related to listener behaviors during the speaker's turn suggests that address prediction should not be treated solely as a problem at turn completion, but rather as something that should be inferred online during ongoing speech. In the future, it will be important to develop frameworks that estimate addressivity in real time while jointly modeling listener behaviors such as gaze and backchannels, as well as turn-taking.

Reconsidered in light of the turn-taking model of Sacks et al. (1974), the present results suggest that the current speaker may be understood not as discretely selecting one next speaker or addressee, but as continuously distributing degrees of addressivity across participants through multimodal interactional resources (Kadota et al., 2024), thereby using listener-wise distributions of addressivity to fine-tune both next-speaker relevance and the uptake of the ongoing action.

Future work should examine whether these findings generalize across languages, domains, and group sizes, and should develop more refined methods for constructing and validating continuous representations of addressivity. Given the links observed here between addressivity and listener gaze and backchannels, future systems may benefit from estimating addressivity online while jointly modeling multiple listener behaviors.

Acknowledgments

This work was supported by JST Moonshot R&D JPMJPS2011 and JST PRESTO JPMJPR24I4.

Limitations

The data are limited to Japanese triadic discussions, and gaze behavior and backchannel production may vary across languages, cultures, settings, and participant characteristics. Although address level was estimated as a latent continuous variable, it was inferred from only three categorical annotations per turn; the behavioral analyses also used posterior medians rather than propagating full posterior uncertainty. Finally, annotators were allowed to use gaze information when judging addressees. Because this primarily concerned the speaker's gaze, whereas our analysis focused on listener gaze toward the current speaker, the gaze analysis does not simply reuse the same behavioral signal, but it should still be interpreted as evidence of consistency rather than as fully independent validation.

References

Oleg Akhtiamov, Maxim Sidorov, Alexey A Karpov, and Wolfgang Minker. 2017. Speech and text analysis for multimodal addressee detection in human-human-computer interaction. In *Interspeech*, pages 2521–2525.

Naoya Baba, Hung-Hsuan Huang, and Yukiko I Nakano. 2011. Identifying utterances addressed to an agent in multiparty human–agent conversations. In *International Workshop on Intelligent Virtual Agents*, pages 255–261. Springer.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3682–3692.

Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, and Tatsuya Kawahara. 2025. An LLM benchmark for addressee recognition in multi-modal multi-party dialogue. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 330–334.

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2006. Addressee identification in face-to-face meetings. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 169–176.

Keisuke Kadota, Seima Oyama, and Yasuharu Den. 2024. Annotation of addressing behavior in multiparty conversation. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Thao Minh Le, Nobuyuki Shimizu, Takashi Miyazaki, and Koichi Shinoda. 2018. Deep learning based multi-modal addressee recognition in visual scenes with utterances. *arXiv preprint arXiv:1809.04288*.

Taiga Mori, Koji Inoue, Divesh Lala, Keiko Ochi, and Tatsuya Kawahara. 2026. Analysing next speaker prediction in multi-party conversation using multimodal large language models. In *Proceedings of the 16th International Workshop on Spoken Dialogue System Technology*, pages 83–94.

Yukiko I Nakano, Naoya Baba, Hung-Hsuan Huang, and Yuki Hayashi. 2013. Implementation and evaluation of a multimodal addressee identification mechanism for multiparty conversation systems. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 35–42.

Harm op den Akker and Rieks op den Akker. 2009. Are you being addressed? - real-time addressee detection to support remote participants in hybrid meetings. In *Proceedings of the SIGDIAL 2009 Conference*, pages 21–28, London, UK. Association for Computational Linguistics.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143.

Nicolò Penzo, Maryam Sajedinia, Bruno Lepri, Sara Tonelli, and Marco Guerini. 2024. Do LLMs suffer from multi-party hangover? a diagnostic approach to addressee recognition and response selection in conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11210–11233.

Suman V Ravuri and Andreas Stolcke. 2014. Neural network models for lexical addressee detection. In *Interspeech*, pages 298–302.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Motoki Sato, Hiroki Ouchi, and Yuta Tsuboi. 2018. Addressee and response selection for multilingual conversation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3631–3644.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Larry Heck. 2012. Learning when to listen: Detecting system-addressed speech in human-human-computer dialog. In *Interspeech*, pages 334–337.

Michael Smithson and Jay Verkuilen. 2006. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54.

TJ Tsai, Andreas Stolcke, and Malcolm Slaney. 2015. A study of multimodal addressee detection in human-human-computer interaction. *IEEE Transactions on Multimedia*, 17(9):1550–1561.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2018. Addressee and response selection in multi-party conversations with speaker interaction RNNs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Pengcheng Zhu, Wei Zhou, Kuncai Zhang, Yuankai Ma, and Haiqing Chen. 2023. Robust learning for multiparty addressee recognition with discrete addressee codebook. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 571–578.

A Detailed Results for the Next-Speaker Models

This appendix reports the full posterior summaries for Models 1a and 1b for next-speaker selection. Both models were fit to 5,820 observations using four chains with 2,000 iterations per chain, including 1,000 warm-up iterations, yielding 4,000 post-warm-up draws in total.

Table 6 summarizes the fixed effects. In Model 1a, the coefficient for *addressed* is positive, indicating that listeners categorized as addressed were more likely to become the next speaker than non-addressed listeners. In Model 1b, the coefficient for *addressLevel* is positive, indicating that the probability of next-speaker selection increases as address level increases.

Model	Param.	Est.	SE	l-95% CI	u-95% CI
(1a)	Intercept*	-0.49	0.13	-0.76	-0.23
(1a)	addressed*	0.72	0.06	0.59	0.84
(1b)	Intercept*	-0.75	0.14	-1.03	-0.49
(1b)	addressLevel*	1.06	0.09	0.89	1.23

Table 6: Posterior summaries of the fixed effects for the next-speaker models. An asterisk (*) indicates that the 95% credible interval does not include zero.

Model	Random effect	Est.	SE	l-95% CI	u-95% CI
(1a)	listenerID	0.65	0.09	0.49	0.84
(1a)	sessionID	0.03	0.02	0.00	0.09
(1a)	speakerID	0.30	0.06	0.19	0.44
(1a)	turnID	0.02	0.02	0.00	0.06
(1b)	listenerID	0.64	0.09	0.49	0.85
(1b)	sessionID	0.03	0.02	0.00	0.08
(1b)	speakerID	0.29	0.06	0.18	0.43
(1b)	turnID	0.02	0.02	0.00	0.06

Table 7: Posterior summaries of the random-effect standard deviations for the next-speaker models.

Table 7 reports the posterior summaries of the random-effect standard deviations. In both models, listener-level variation was the largest source of heterogeneity, followed by speaker-level variation, whereas session- and turn-level variation were comparatively small.

All parameters showed satisfactory convergence diagnostics, with $\hat{R} = 1.00$ throughout. Bulk ESS and Tail ESS were also sufficiently large for all reported parameters.

B Detailed Results for the Listener-Gaze Models

This appendix reports the full posterior summaries for Models 2a and 2b for listener gaze. Both models were fit to 5,744 observations using four chains with 2,000 iterations per chain, including 1,000 warm-up iterations, yielding 4,000 post-warm-up draws in total.

Table 8 summarizes the fixed effects and the beta precision parameter. In Model 2a, the coefficient for *addressed* is positive, indicating that listeners categorized as addressed showed higher gaze proportions toward the current speaker. In Model 2b, the coefficient for *addressLevel* is also positive, indicating that expected gaze proportion increases as address level increases.

Table 9 reports the posterior summaries of the random-effect standard deviations. In both models, turn-level variation was the largest source of heterogeneity, followed by listener- and speaker-level variation, whereas session-level variation was comparatively smaller.

Model	Param.	Est.	SE	l-95% CI	u-95% CI
(2a)	Intercept*	-0.78	0.11	-0.99	-0.57
(2a)	addressed*	0.77	0.04	0.69	0.85
(2a)	ϕ*	0.83	0.02	0.79	0.88
(2b)	Intercept*	-1.06	0.11	-1.29	-0.84
(2b)	addressLevel*	1.15	0.05	1.04	1.26
(2b)	ϕ*	0.84	0.02	0.79	0.89

Table 8: Posterior summaries of the fixed effects and beta precision parameter for the listener-gaze models. An asterisk (*) indicates that the 95% credible interval does not include zero.

Model	Random effect	Est.	SE	l-95% CI	u-95% CI
(2a)	listenerID	0.47	0.07	0.35	0.63
(2a)	sessionID	0.22	0.05	0.14	0.32
(2a)	speakerID	0.35	0.06	0.25	0.49
(2a)	turnID	0.58	0.04	0.50	0.65
(2b)	listenerID	0.47	0.07	0.35	0.63
(2b)	sessionID	0.21	0.04	0.14	0.31
(2b)	speakerID	0.34	0.06	0.24	0.48
(2b)	turnID	0.57	0.04	0.49	0.64

Table 9: Posterior summaries of the random-effect standard deviations for the listener-gaze models.

Convergence diagnostics were satisfactory overall. Most reported parameters had $\hat{R}$ values of 1.00, while the turn-level standard deviation and ϕ in Model 2b had $\hat{R} = 1.01$, which still indicates acceptable convergence. Bulk ESS and Tail ESS were also sufficiently large for the reported parameters.

C Detailed Results for the Backchannel Models

This appendix reports the full posterior summaries for Models 3a and 3b for backchannel production. Both models were fit to 4,920 observations using four chains with 2,000 iterations per chain, including 1,000 warm-up iterations, yielding 4,000 post-warm-up draws in total.

Table 10 summarizes the fixed effects. In Model 3a, the coefficient for *addressed* is positive, indicating that listeners categorized as addressed produced more backchannels than non-addressed listeners. In Model 3b, the coefficient for *addressLevel* is positive, indicating that the expected backchannel rate increases as address level increases.

Table 11 reports the posterior summaries of the random-effect standard deviations. In both models, listener-level variation was the largest source of heterogeneity, whereas session-, speaker-, and turn-level variation were comparatively small.

Convergence diagnostics were satisfactory overall. Most reported parameters had $\hat{R}$ values of 1.00, while a small number of parameters had $\hat{R} = 1.01$, which still indicates acceptable convergence. Bulk

Model	Param.	Est.	SE	l-95% CI	u-95% CI
(3a)	Intercept*	-2.37	0.10	-2.58	-2.17
(3a)	addressed*	0.18	0.05	0.08	0.28
(3b)	Intercept*	-2.48	0.11	-2.69	-2.26
(3b)	addressLevel*	0.31	0.07	0.17	0.45

Table 10: Posterior summaries of the fixed effects for the backchannel models. An asterisk (*) indicates that the 95% credible interval does not include zero.

Model	Random effect	Est.	SE	l-95% CI	u-95% CI
(3a)	listenerID	0.51	0.07	0.38	0.67
(3a)	sessionID	0.09	0.04	0.01	0.18
(3a)	speakerID	0.12	0.05	0.03	0.22
(3a)	turnID	0.05	0.03	0.00	0.12
(3b)	listenerID	0.50	0.07	0.38	0.66
(3b)	sessionID	0.08	0.04	0.01	0.17
(3b)	speakerID	0.12	0.05	0.02	0.23
(3b)	turnID	0.04	0.03	0.00	0.12

Table 11: Posterior summaries of the random-effect standard deviations for the backchannel models.

ESS and Tail ESS were also sufficiently large for the reported parameters.