Rethinking Binary Evaluation of Turn-Taking under Inherent Ambiguity

Yunosuke Kubo, Kenta Yamamoto, Ryu Takeda and Kazunori Komatani

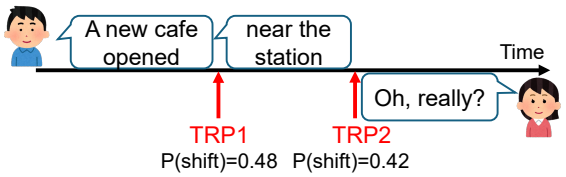
SANKEN, University of Osaka
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
Correspondence: komatani@sanken.osaka-u.ac.jp

Abstract

Turn-taking prediction models output probabilities of turn shifts, yet they are typically evaluated by thresholding these probabilities into binary decisions and comparing them against corpus-observed labels. This practice implicitly treats corpus-observed turn shifts as definitive ground truth, even though under inherent turn-taking ambiguity they reflect one realized interactional outcome among multiple plausible outcomes, rather than a uniquely correct binary label. We argue that binary evaluation is a practical simplification rather than a theoretical necessity. Instead, predicted probabilities should be evaluated at the distributional level without being reduced to binary decisions. To this end, we propose a distribution-based evaluation framework that compares model output distributions with reference distributions and measures their divergence using the Wasserstein distance. We further show how discrepancies between model predictions and corpus-observed turn shifts can be used as a basis for training-data refinement. Experiments on Japanese conversational data, using linguistic information alone, showed that the proposed refinement reduced distributional divergence, indicating better alignment between predicted probabilities and the reference distributions. The refinement also improved balanced accuracy in a supplementary binary evaluation.

1 Introduction

Spoken dialogue systems have become increasingly prevalent, and recent advances in large language models (LLMs) have significantly improved the quality of system responses. While much progress has been made in modeling what to say, determining when to speak remains a fundamental challenge in spoken interaction. Compared with text-based interaction, spoken interaction provides fewer explicit cues to user turn completion, so systems must predict turn transitions and initiate responses at appropriate moments. Accurate turn-taking prediction is essential to avoid response overlaps and unnatural gaps, both of which degrade interaction quality.



Issue 1:
Thresholding predicted probabilities into hard labels
Issue 2:
Treating the corpus-observed outcome as ground truth

Figure 1: Two issues in binary evaluation at TRPs.

A wide range of approaches has been proposed for turn-taking prediction, including silence-based thresholding and machine learning models leveraging acoustic, linguistic, and multimodal features. Recent probabilistic models estimate the likelihood of turn shift at each time step. Despite producing probabilistic outputs, these models are typically evaluated through binary thresholding, where predicted probabilities are converted into discrete shift/non-shift decisions.

Figure 1 highlights two issues in binary evaluation at transition relevance places (TRPs). First, probabilistic model outputs are collapsed into hard labels. Second, the corpus-observed outcome is treated as absolute ground truth, even though under turn-taking ambiguity it is one realized outcome among several plausible alternatives. In natural interaction, such corpus observations do not uniquely determine a correct binary label, and treating them as definitive labels overlooks the inherent ambiguity of turn-taking.

In this paper, we revisit binary evaluation for turn-taking prediction under inherent ambiguity, focusing on prediction based solely on linguistic information. We propose a distribution-based eval-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 13–24
August 2–5, 2026.

uation framework for assessing turn-shift probabilities directly, and further show how the same ambiguity-aware perspective can be used to refine training data. Experiments show that this refinement reduces distributional divergence under the proposed evaluation framework and also improves balanced accuracy in a supplementary binary evaluation.

This paper makes the following contributions:

- We argue that corpus-observed turn shifts should be regarded as realized outcomes rather than absolute ground truth, highlighting the inherent ambiguity of turn-taking.
- From this perspective, we propose a distribution-based evaluation framework that assesses predicted turn-shift probabilities without collapsing them into binary labels.
- We demonstrate that the same ambiguity-aware perspective can be used diagnostically to identify prediction-corpus discrepancies and guide refinement of training data.
- We empirically validate the proposed perspective by showing that training-data refinement reduces distributional divergence under the proposed evaluation framework and improves balanced accuracy in a supplementary binary evaluation.

2 Related Work

2.1 Turn-Taking Theory and Ambiguity

Turn-taking in human conversation has been systematically described in conversation analysis. Sacks et al. (1974) demonstrated that turn-taking follows structured rules rather than occurring randomly. Utterances are organized into turn-constructional units (TCUs), and positions at which a TCU may reach possible completion are referred to as transition relevance places (TRPs). Importantly, a TRP marks a position where speaker change may occur, but does not guarantee that it will.

When multiple TRPs appear within a turn, listeners decide at each point whether to start speaking. Consequently, an observed turn shift in a corpus represents only one realized outcome among multiple possible continuations. This suggests that turn-taking is more naturally treated as a probabilistic rather than binary process.

Psycholinguistic studies also indicate that listeners do not wait until utterance completion to prepare their response, but instead process comprehension and prediction in parallel (Garrod and Pickering, 2015). Prosodic cues, utterance length, and syntactic and semantic completeness have been shown to influence turn-shift timing (Gravano and Hirschberg, 2011). De Ruiter et al. (2006) demonstrated that linguistic information plays a central role in projecting utterance completion. These findings support the view that turn-taking involves anticipatory probabilistic inference rather than a deterministic binary process.

2.2 Turn-Taking Prediction

A wide range of approaches has been proposed for turn-taking prediction. Early implementations primarily relied on silence duration thresholds to detect turn completion. However, setting appropriate thresholds that simultaneously suppress false cut-ins and response latency is challenging. Raux and Eskenazi (2008) proposed dynamically adapting silence thresholds based on dialogue context and interactional features to address this issue.

More recent approaches leverage acoustic, linguistic, and multimodal features. Language-model-based approaches such as TurnGPT (Ekstedt and Skantze, 2020) estimate the probability of turn shift at each input position. Related studies explore projection-based methods (Ekstedt and Skantze, 2021), response-conditioned modeling (Jiang et al., 2023), negative sampling strategies (Muromachi and Kano, 2023), and in-context learning (Umair et al., 2024).

Voice-based approaches such as VAP (Ekstedt and Skantze, 2022) directly predict future voice activity and have demonstrated real-time applicability, with subsequent validation (Inoue et al., 2024) and extension (Inoue et al., 2025). Attempts to integrate speech and linguistic information have also been explored (Hara et al., 2019; Li et al., 2022; Sakuma et al., 2022; Wang et al., 2024; Skantze and Irfan, 2025), though temporal alignment and modality fusion remain challenging.

Despite differences, many contemporary models output turn-shift probabilities.

2.3 Evaluation Practices

Despite their probabilistic nature, turn-taking prediction models are typically evaluated through binary classification: corpus-observed turn shifts are treated as ground truth, and predicted probabili-

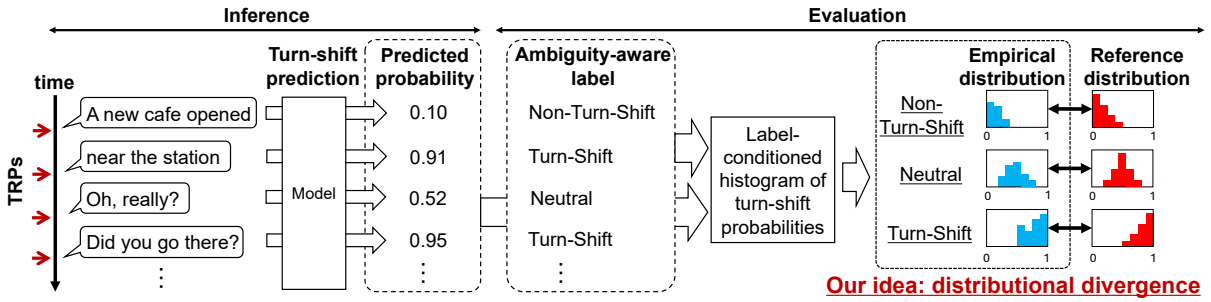


Figure 2: Overview of distribution-based evaluation of turn-shift probabilities

ties are thresholded to compute accuracy or related metrics. Although some studies conceptually treat turn-taking opportunities at TRPs as probabilistic, evaluation procedures typically reduce predicted probabilities to binary decisions.

Direct annotation of TRPs is costly and rarely available at scale. Some studies have constructed datasets that approximate TRP labels (Hara et al., 2019) or are based on behavioral evidence from participants (Umair et al., 2024).

Although the non-uniqueness of observed turn-taking outcomes is related to a general issue in language modeling, standard metrics such as perplexity primarily evaluate probability assigned to corpus-observed sequences. By contrast, the proposed evaluation examines label-conditioned distributions of turn-shift probabilities with respect to ambiguity-aware reference distributions.

In turn-taking prediction, evaluation remains largely tied to corpus-observed shifts, which may reflect only one realized outcome among multiple plausible turn-taking outcomes. This gap between theoretical ambiguity and binary evaluation motivates the present study.

3 Distribution-Based Evaluation for Turn-Shift Probabilities

In contrast to prior work that reduces probabilistic outputs to binary decisions, we propose a distribution-based evaluation framework for turn-shift probabilities. Our objective is to assess whether predicted probability mass reflects the inherent ambiguity of turn boundaries, rather than merely matching corpus-observed binary shift annotations.

We treat turn-taking as a probabilistic phenomenon grounded in the non-uniqueness of turn-taking outcomes at transition relevance places (TRPs), where both continuation and transition may be plausible. Instead of thresholding model outputs, we compare empirical distributions of predicted probabilities with reference distributions. The divergence between these distributions serves as a measure of how well predicted probabilities align with the reference distributions.

The proposed distribution-based evaluation consists of three steps: (i) assigning three ambiguity-aware labels to evaluation instances, (ii) constructing empirical probability distributions conditioned on these labels, and (iii) measuring their divergence from predefined reference distributions. Figure 2 summarizes the evaluation procedure.

3.1 Ambiguity-Aware Evaluation Labels

Because our study focuses on prediction based solely on linguistic information, evaluation is also conducted using text-only criteria. We do not assume that syntactic and semantic completeness fully determines turn-taking in natural conversation; rather, these notions serve as operational criteria for what can be inferred from linguistic information alone. For each candidate boundary in the evaluation data, we assign one of three labels reflecting the degree of turn-taking ambiguity:

- **Non-Turn-Shift (NTS)**: The utterance is syntactically or semantically incomplete, and a turn shift is not linguistically projected at this point.
- **Turn-Shift (TS)**: The utterance is syntactically and semantically sufficiently complete, and a turn shift is linguistically plausible.
- **Neutral (Neu)**: The completeness of the utterance cannot be reliably determined from linguistic information alone, and both continuation and transition are plausible.

The introduction of the *Neutral* category explicitly acknowledges that corpus-observed turn shifts are realized outcomes rather than absolute

ground truth. It captures cases where linguistic cues alone are insufficient and additional modalities (e.g., prosody or timing) may be required to interpret their turn-taking status, thereby avoiding forced binary decisions for cases that are uncertain under text-only criteria.

3.2 Construction of Empirical Output Distributions

To evaluate probability estimates without thresholding, we analyze the label composition within each predicted-probability bin.

Let the evaluation dataset be $D = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes an input instance, $y_i \in L$ is the ambiguity-aware label assigned as described in Section 3.1, and $p_i \in [0, 1]$ is the predicted probability of a turn shift output by the model.

We partition the probability interval $[0, 1]$ into M bins $B_1, \ldots, B_M$, and define $S_k = \{i \mid p_i \in B_k\}$ as the set of instances whose predicted probabilities fall into bin B_k. In all experiments, we used $M = 25$ equal-width bins over $[0, 1]$, resulting in a bin width of 0.04. The same binning was used to construct both the empirical and reference distributions. For each label $l \in L$, we define its proportion among the instances in bin B_k as

$$v_{k,l} = \frac{|\{i \in S_k \mid y_i = l\}|}{|S_k|}. \tag{1}$$

We set $v_{k,l} = 0$ whenever $|S_k| = 0$.

The vector $\{v_{k,l}\}_{l \in L}$ for a fixed bin B_k represents the label composition within that bin. Collecting these vectors across bins yields an empirical histogram showing how ambiguity-aware labels are distributed over the predicted-probability axis. This representation allows us to examine, for each probability range, what proportion of instances are labeled as *Non-Turn-Shift*, *Neutral*, or *Turn-Shift*, thereby revealing whether higher predicted probabilities are more often associated with instances labeled as *Turn-Shift*.

Unlike binary accuracy metrics, this distributional view preserves graded information and enables a finer-grained assessment of probabilistic behavior.

3.3 Design of Reference Distributions

The reference distributions specify expected probability patterns for the ambiguity-aware labels along the probability axis. These distributions serve as evaluation references rather than as attempts to estimate true human response distributions. Instances

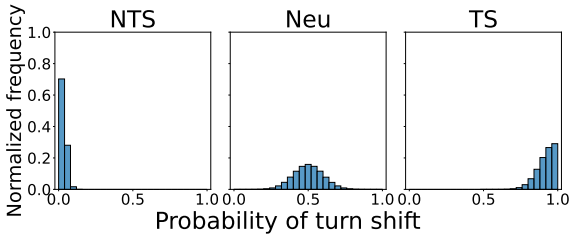


Figure 3: Gaussian-shaped reference distributions for NTS, Neu, and TS.

labeled as *Non-Turn-Shift* are expected to concentrate near 0, those labeled as *Turn-Shift* near 1, and those labeled as *Neutral* around the midpoint, reflecting their relative appropriateness for a turn shift. To operationalize this intuition, we adopt Gaussian-shaped reference distributions for the three labels. Because this design is hand-specified, we examine robustness to alternative reference distributions in Appendix A.

For each label $l \in \{\mathrm{NTS}, \mathrm{Neu}, \mathrm{TS}\}$, we define a Gaussian reference distribution with mean μ_l and standard deviation σ_l as follows:

$$(\mu_l, \sigma_l) = \begin{cases} (0.02, 0.027) & \text{if } l = \mathrm{NTS}, \\ (0.50, 0.100) & \text{if } l = \mathrm{Neu}, \\ (0.98, 0.093) & \text{if } l = \mathrm{TS}. \end{cases}$$

Figure 3 illustrates the resulting reference distributions.

3.4 Distributional Distance (Wasserstein)

For each label, we quantify the discrepancy between the empirical distribution and its corresponding reference distribution. Smaller values indicate closer alignment between predicted probabilities and the corresponding reference distribution.

We adopt the Wasserstein distance (optimal transport) as the divergence measure, which quantifies the cost of transporting probability mass (Villani, 2009). Because it incorporates the geometry of the probability axis, it distinguishes between small and large displacements of probability mass.

This property is well suited to turn-shift probabilities. Moving mass from 0.49 to 0.51 is treated as much smaller than moving it from 0.01 to 0.99. Bin centers are treated as support coordinates and bin values as probability masses in the optimal transport computation.

4 Data Refinement under Turn-Taking Ambiguity

Building on the premise that corpus-observed turn shifts represent realized outcomes rather than absolute ground truth, we explore a second practical implication of the ambiguity-aware perspective introduced in Section 3. While the previous section reconsidered evaluation under inherent ambiguity, this section examines how the same perspective can guide training-data refinement.

In spoken dialogue corpora, observed turn shifts are shaped not only by linguistic completeness but also by prosody, timing, gaze, and other interactional factors. As a result, some corpus-observed shifts may be difficult to explain from linguistic information alone. When a text-based model is trained on such data, prediction-corpus discrepancies can highlight instances where the corpus-observed outcome is not strongly supported by linguistic patterns learned from the broader training data. The refinement procedures explored in this section use these discrepancies as diagnostic signals. They do not assume that the discrepant instances are incorrect; instead, they examine how treating such instances differently affects the resulting probabilistic turn-shift predictions.

4.1 Identification of Discrepant Instances

We first obtain predicted probabilities p_i for each instance i using an initial model fine-tuned on the full training dataset D. Let $y_i \in \{0, 1\}$ denote the corpus-based binary label, where 1 indicates a turn shift and 0 indicates no shift.

Although our primary evaluation is distribution-based, we temporarily introduce a threshold θ to derive a binary prediction:

$$\hat{y}_i = \begin{cases} 1 & (p_i \geq \theta) \\ 0 & (p_i < \theta) \end{cases}. \tag{2}$$

This thresholded decision is used solely to identify mismatches between model predictions and corpus-based binary labels.

We group these discrepant cases into false negatives (FN) and false positives (FP), defined as

$$S_{FN} = \{i \mid y_i = 1, \hat{y}_i = 0\},$$
$$S_{FP} = \{i \mid y_i = 0, \hat{y}_i = 1\}.$$

These discrepant instances are treated as candidates for data refinement.

4.2 Refinement Operations

The discrepant cases identified above may arise from different sources. Some FN instances may involve turn shifts that are difficult to explain from linguistic information alone, whereas some FP instances may correspond to linguistically plausible TRPs at which the speaker nevertheless chose to continue. These considerations motivate the refinement strategies described below.

We consider four refinement strategies based on the discrepant instance sets defined above, each producing a refined training dataset D':

1. **FN Removal**: $D' = D \setminus S_{FN}$. This operation removes instances for which the corpus indicates a turn shift but the model assigns a low shift probability.

2. **FP Removal**: $D' = D \setminus S_{FP}$. This operation removes instances for which the model assigns a high shift probability despite the absence of an observed turn shift.

3. **FN+FP Removal**: $D' = D \setminus (S_{FN} \cup S_{FP})$. This operation removes all instances for which model predictions and corpus-based binary labels diverge.

4. **FN Relabeling**: The refined training dataset D' is constructed with training labels y'_i, where $y'_i = 0$ for $i \in S_{FN}$ and $y'_i = y_i$ otherwise. This operation treats the identified FN instances as non-shift cases during training.

We do not consider FP relabeling, as assigning turn-shift labels to mid-utterance positions may create linguistically implausible transition points.

5 Experimental Evaluation

In this section, we evaluate the effect of training-data refinement. We first describe the data preprocessing and annotation procedure, and then compare a baseline model with refinement variants using the proposed distribution-based metric. We also report supplementary binary classification metrics for comparison with conventional practice.

5.1 Data and Preprocessing

5.1.1 Dataset

We used the Corpus of Everyday Japanese Conversation (CEJC) (Koiso et al., 2022), a large-scale corpus of spontaneous everyday Japanese conversation. It contains 577 dialogues (approximately

	Train	Valid	Eval
Dialogues	520	29	–*
Turns	177,180	11,101	500

* Eval: 500 turns sampled from the remaining dialogues.

Table 1: Data statistics.

Label	Count
Turn-Shift	249
Non-Turn-Shift	88
Neutral	163

Table 2: Label distribution in the evaluation set.

Speaker	Utterance
A	Yeah, that's true.
B	Ah, I think it just keeps going.
C	Oh, really?
D	The experience, um, well...
B	**I mean, "look,"...**

Figure 4: An example annotated as *Neutral*, with the final utterance as the target of annotation.

200 hours of recordings) involving 1,675 speakers of diverse ages and genders, and includes audio, video, and manually transcribed text.

Since our goal is to predict turn-shift probabilities from linguistic information alone, we used the manually transcribed text with speaker identifiers. The corpus includes both dyadic and multi-party conversations. Because it consists of spontaneous everyday dialogue, utterances are often casual and disfluent, and turn shifts do not always coincide with clearly complete linguistic boundaries. In our experiments, a turn shift was defined as a change in speaker identity, regardless of which speaker took the next turn.

Because CEJC was not originally designed for turn-shift prediction, we applied preprocessing. We removed transcription tags that would not be available in real-world automatic speech recognition outputs (e.g., laughter markers, prosodic annotations such as rising intonation, and uncertain segments), merged consecutive long-utterance units (LUUs) by the same speaker into a single turn, and excluded heavily overlapping cases (see Appendix B for details). After preprocessing, each turn corresponded to consecutive LUUs produced by a single speaker.

5.1.2 Data Splits and Annotation

We split the 577 dialogues in CEJC into training, validation, and evaluation sets at the dialogue level. Ninety percent (520 dialogues) were used for training and 5% (29 dialogues) for validation. The validation set was used for early stopping during model training. From the remaining 28 dialogues, we extracted 500 turns to construct the evaluation set for manual annotation. Table 1 summarizes the resulting data sizes.

For the evaluation set, one of the authors manually annotated each instance using the text transcripts with speaker identifiers. The annotation target for each instance was the end of a sampled corpus turn. During annotation, the annotator considered the target turn together with the preceding four turns as dialogue history. Following Section 3.1, we adopted the three ambiguity-aware labels and each instance was labeled as *Turn-Shift*, *Non-Turn-Shift*, or *Neutral*. These labels indicate the linguistic plausibility of a turn shift, rather than whether a speaker change was observed in the corpus; therefore, even corpus-observed turn boundaries may be labeled as *Non-Turn-Shift* or *Neutral*. The distribution of the annotated labels is shown in Table 2. Among the 500 annotated turns, 249 were labeled as *Turn-Shift*, 88 as *Non-Turn-Shift*, and 163 as *Neutral*. The relatively high proportion of *Turn-Shift* labels reflects the fact that annotation was performed at actual turn boundaries in the corpus.

Figure 4 shows an example labeled as *Neutral*, translated from Japanese. The final utterance can be interpreted either as projecting continuation by the same speaker or as sufficiently complete to allow speaker change. Because both interpretations are plausible, the utterance was labeled as *Neutral*.

5.2 Experimental Setup

We fine-tuned `japanese-gpt-1b`, a language model provided by rinna Co., Ltd.[1], to predict turn-shift probabilities, following the approach of TurnGPT (Ekstedt and Skantze, 2020). Specifically, we introduced a special token ⟨ts⟩ representing a potential turn shift and inserted it at the end of each turn in the training data. Given the preceding dialogue context, the model learned to predict the probability of generating the ⟨ts⟩ token, which we interpret as the probability of a turn shift at that point. To incorporate conversational context, each input consisted of the current turn and its preceding

[1] https://huggingface.co/rinna/japanese-gpt-1b

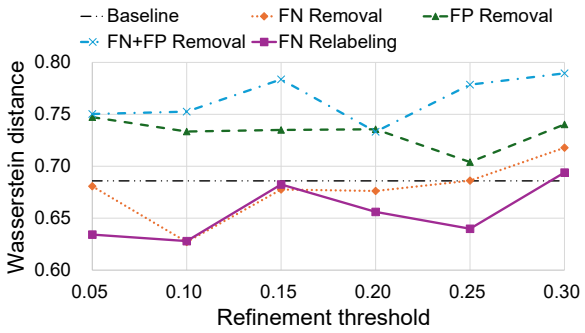


Figure 5: Performance across refinement thresholds for different refinement strategies.

turns. All experiments used a four-turn dialogue history, selected based on preliminary experiments. Additional training details, including the full training configuration, are provided in Appendix B.

Following the refinement framework introduced in Section 4.2, we evaluated four refinement conditions: FN Removal, FP Removal, FN+FP Removal, and FN Relabeling. In addition to these conditions, we included a baseline model trained on the original data without refinement.

Discrepant instances were detected using the turn-shift probabilities predicted by the baseline model. Unless otherwise stated, we set $\theta = 0.10$ in Eq. 2 for discrepant-instance detection.

We used the Gaussian-shaped reference distributions shown in Figure 3, where the *Neutral* distribution is centered at 0.5. Robustness to alternative reference distributions is reported in Appendix A.

5.3 Distribution-Based Evaluation

We evaluated model outputs on CEJC using the proposed distribution-based metric. We first analyzed how different refinement strategies affected the predicted probability distributions under this metric, and then examined label-wise changes in the output distributions.

5.3.1 Effect of Refinement

Figure 5 shows the summed Wasserstein distance across refinement thresholds for the four refinement strategies. The vertical axis represents the summed Wasserstein distance over all labels, and the horizontal axis represents the refinement threshold. Lower values indicate better performance. The baseline represents the model trained on the original training data without refinement.

The results show that the FN-based strategies reduced the summed Wasserstein distance relative to the baseline, whereas the FP-based strategies did not. For FN Removal, the Wasserstein distance was smaller than the baseline distance for thresholds between 0.05 and 0.20, while for FN Relabeling, it was smaller for thresholds between 0.05 and 0.25. In both strategies, the distance reached its minimum at a threshold of 0.10.

This difference suggests that FN and FP discrepancies should not be treated symmetrically in this setting. FN instances correspond to corpus-observed turn shifts that the text-based model did not strongly support. Such cases may include shifts driven by prosody, timing, overlap, or other interactional factors that are not well captured by linguistic information alone. FN-based refinement reduces the influence of treating such corpus-observed shifts as fixed turn-shift labels during training, which can make the resulting probability distributions more consistent with the reference distributions. By contrast, FP-based refinement removes corpus-observed non-shift cases at points where turn shifts appear linguistically plausible. Although this operation is a reasonable refinement strategy, the results show that it did not improve alignment with the reference distributions.

These results indicate that, among the refinement strategies examined here, the FN-based strategies with a threshold of 0.10 yielded the lowest Wasserstein distances under the proposed distribution-based evaluation.

5.3.2 Label-wise Analysis

We next analyzed how the model output distributions changed for each label after refinement. Figure 6 shows the relative frequency of the three labels within each probability bin. Each bin represents the proportions of *Non-Turn-Shift*, *Neutral*, and *Turn-Shift* labels among instances whose predicted turn-shift probabilities fall in that range. The figure compares the baseline model with the FN Removal and FN Relabeling models at a refinement threshold of 0.10.

For the *Turn-Shift* label, both FN Removal and FN Relabeling produced output probabilities that were more concentrated near 1.0 than the baseline. This shift toward the high-probability region is consistent with the reference distribution, which assigns higher density near 1.0. Table 3 further reports the label-wise Wasserstein distances to the reference distributions. The reduction was particularly large for the *Turn-Shift* label. The distance decreased from 0.376 in the baseline model to 0.319 for FN Relabeling and 0.326 for FN Removal.

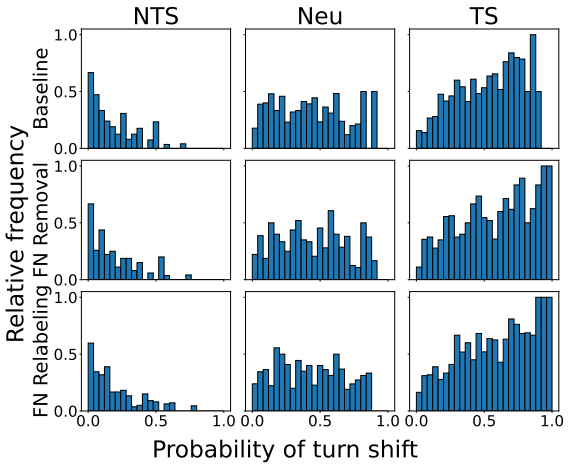


Figure 6: Relative frequency of the three labels across probability bins.

Label	Baseline	FN Relabeling	FN Removal
NTS	0.158	0.166	0.160
TS	0.376	**0.319**	**0.326**
Neu	0.153	0.143	0.141
Total	0.686	0.628	**0.627**

Table 3: Label-wise Wasserstein distance to the reference distribution for the baseline and FN-based refinement strategies.

This pattern is consistent with the possibility that some false-negative instances in the original training data involved observed turn shifts that were weakly supported by linguistic information alone. Because CEJC contains casual and multi-party conversations, such cases may arise from overlaps, interruptions, or other conversational dynamics. Although such transitions can be interactionally natural, they may be difficult to predict from linguistic information alone and may therefore appear as false negatives in text-based models. At the same time, we do not assume that all false-negative instances arise from such conversational dynamics; some may also reflect limitations of the model, the available features, or the amount of training data. By changing the treatment of identified false-negative instances during training, FN-based refinement produced a *Turn-Shift* distribution closer to the reference distribution.

5.4 Binary Evaluation

For comparison with conventional practice, we also report a supplementary binary classification evaluation. *Neutral* instances were excluded from this comparison because they do not correspond directly to either class in a conventional binary turn-shift classification task. Accordingly, this evaluation does not capture the ambiguity addressed by the proposed distribution-based metric.

Model	BAcc
Baseline	0.777
FN Removal	0.788
FN Relabeling	**0.799**

Table 4: Binary classification performance (BAcc). Results are averaged over two-fold cross-validation.

Based on the distribution-based evaluation above, we compared the baseline model with the two FN-based refinement strategies, FN Removal and FN Relabeling, using a refinement threshold of 0.10. Details of the binary evaluation procedure are provided in Appendix C.

Table 4 shows the results averaged over the two cross-validation trials. The baseline model achieved a BAcc of 0.777, while FN Removal and FN Relabeling achieved 0.788 and 0.799, respectively. These results suggest that the improvement trend observed in the distribution-based evaluation also appeared in the binary evaluation.

5.5 Implications for Turn-taking Modeling

Our results suggest that the refinement operations change how corpus-observed turn shifts are used as training signals. The resulting probability outputs became more consistent with reference distributions designed to reflect linguistic plausibility. In this sense, their effect should not be interpreted simply as noise removal.

Conventional binary evaluation can compare models after assigning each instance to one of two classes, but it cannot directly assess how model probabilities behave for *Neutral* or otherwise ambiguous instances. Explicitly representing such instances allows cases that are difficult to resolve under the available input modality to be examined directly.

Turn-taking opportunities are often inherently ambiguous, and multiple outcomes may be interactionally plausible. In such cases, evaluating models against a single deterministic label may obscure important aspects of model behavior, particularly at ambiguous turn boundaries. These findings suggest that turn-taking models should explicitly account for such ambiguity rather than treating corpus-observed outcomes as uniquely correct targets.

6 Conclusion

This paper argued that turn-taking prediction requires ambiguity-aware evaluation. To address this issue, we proposed a distribution-based evaluation framework that compares predicted turn-shift probabilities with ambiguity-aware reference distributions using the Wasserstein distance. Using this framework, we analyzed the effect of training-data refinement and found that FN-based refinement brought the predicted probability distributions into closer alignment with the reference distributions.

These findings are particularly important in the present setting, where the model predicts turn-taking from linguistic information alone. Although real conversational turn-taking is shaped by multimodal signals such as prosody, timing, and gaze, the present results suggest that even in a text-based setting, corpus-observed outcomes should not be treated as uniquely correct binary targets. The proposed framework provides a way to assess such probabilistic behavior more appropriately.

To support further research on Japanese turn-taking prediction, we make publicly available a TurnGPT model retrained on a version of the training data with personal information removed.[2]

Future work includes the data-driven design of reference distributions, validation across languages and model types, and extension of the framework to multimodal turn-taking prediction. More broadly, the present results support the need for ambiguity-aware evaluation of probabilistic turn-taking predictions.

Limitations

This study has several limitations. First, the evaluation labels were manually annotated by one of the authors, and inter-annotator agreement was not assessed. The annotation was intended to provide a practical operationalization of turn-taking ambiguity for analysis, and its reproducibility should be examined in future work using multiple annotators.

Second, the reference distributions used in this study were hand-specified. They encode simple expected patterns of turn-shift probabilities for *Non-Turn-Shift*, *Neutral*, and *Turn-Shift* labels, rather than empirically estimated human response distributions. Although Appendix A suggests that the main trends are robust to several alternative settings, the design of reference distributions remains an open issue. In particular, Gaussian-shaped distributions are only one possible smooth operationalization, and other bounded or asymmetric distributions may be more appropriate for labels near the edges of the probability scale. Future work should estimate or validate such distributions using human response data or multimodal evidence from spoken interaction.

Third, the present study focuses on turn-taking prediction from linguistic information alone. In natural conversation, turn-taking is also shaped by prosody, timing, gaze, gesture, and other multimodal cues. Our findings should therefore be interpreted within a text-based prediction setting. In embodied or multimodal dialogue settings, where such cues are directly available, the appropriate ambiguity-aware labels and reference distributions may differ from those used here. Future work should investigate whether the same evaluation perspective extends to multimodal turn-taking models.

Fourth, the experiments were conducted on a Japanese spoken-dialogue corpus. While the main contribution of this paper is methodological rather than language-specific, the linguistic and interactional cues associated with turn-taking may differ across languages and conversational settings. Future work should therefore test the proposed framework on additional dialogue corpora in other languages, including English.

Fifth, this study did not directly evaluate downstream dialogue-system performance, such as response latency, interruption rates, user satisfaction, or task success. The proposed framework should therefore be understood as an evaluation and diagnostic framework for probabilistic turn-taking predictions, and future work should examine how this type of evaluation relates to behavior in dialogue systems.

Sixth, the empirical analysis was conducted with a single turn-taking prediction model. The main claim of this paper concerns the evaluation of probabilistic turn-shift predictions rather than the superiority of any particular model. At the same time, the quantitative effects of the proposed refinement procedure may vary across model architectures, scales, and training conditions, and should be examined more broadly in future work.

Acknowledgments

We thank the reviewers for their constructive discussions and valuable feedback. This work was partly

[2] https://github.com/osaka-dialogue/TurnGPT-jp

supported by JSPS KAKENHI Grant Numbers JP22H00536 and JP23K28147, and JST Moonshot R&D Grant Number JPMJMS2011, Japan.

References

Jan Peter De Ruiter, Holger Mitterer, and Nick J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2981–2990.

Erik Ekstedt and Gabriel Skantze. 2021. Projection of turn completion in incremental spoken dialogue systems. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 431–437.

Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised learning of turn-taking events. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 5190–5194.

Simon Garrod and Martin J. Pickering. 2015. The use of content and timing to predict turn transitions. *Frontiers in Psychology*, 6(751):1–12.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Turn-taking prediction based on detection of transition relevance place. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4170–4174.

Koji Inoue, Mikey Elmers, Yahui Fu, Zi Haur Pang, Divesh Lala, Keiko Ochi, and Tatsuya Kawahara. 2025. Prompt-guided turn-taking prediction. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 146–151.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Real-time and continuous turn-taking prediction using Voice Activity Projection. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. 2023. Response-conditioned turn-taking prediction. In *Findings of the Association for Computational Linguistics (ACL)*, pages 12241–12248.

Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. 2022. Design and evaluation of the corpus of everyday Japanese conversation. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 5587–5594.

Siyan Li, Ashwin Paranjape, and Christopher D. Manning. 2022. When can I speak? Predicting initiation points for spoken dialogue agents. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 217–224.

Toshiki Muromachi and Yoshinobu Kano. 2023. Estimation of listening response timing by generative model and parameter control of response substantialness using dynamic-prompt-tune. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2638–2642.

Antoine Raux and Maxine Eskenazi. 2008. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 1–10.

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.

Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. 2022. Response timing estimation for spoken dialog system using dialog act estimation. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4486–4490.

Gabriel Skantze and Bahar Irfan. 2025. Applying general turn-taking models to conversational human-robot interaction. In *Proceedings of the 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 859–868.

Muhammad Umair, Vasanth Sarathy, and Jan Peter De Ruiter. 2024. Large Language Models know what to say but not when to speak. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 15503–15514.

Cédric Villani. 2009. *Optimal Transport: Old and New*. Springer.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-taking and backchannel prediction with acoustic and large language model fusion. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12121–12125.

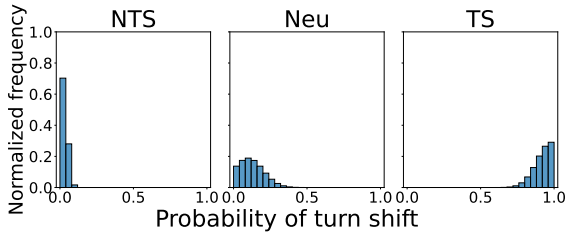


Figure 7: Gaussian-shaped distributions: $[\mu_{\text{NTS}}, \mu_{\text{Neu}}, \mu_{\text{TS}}] = [0.02, 0.10, 0.98]$.

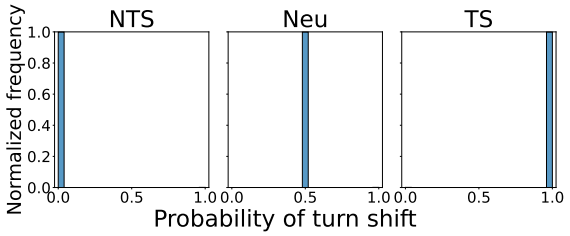


Figure 8: Single-bin distributions: $[\mu_{\text{NTS}}, \mu_{\text{Neu}}, \mu_{\text{TS}}] = [0.0, 0.5, 1.0]$.

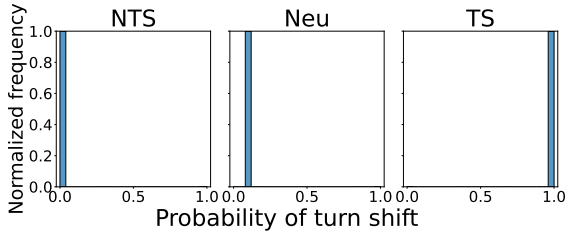


Figure 9: Single-bin distributions: $[\mu_{\text{NTS}}, \mu_{\text{Neu}}, \mu_{\text{TS}}] = [0.0, 0.1, 1.0]$.

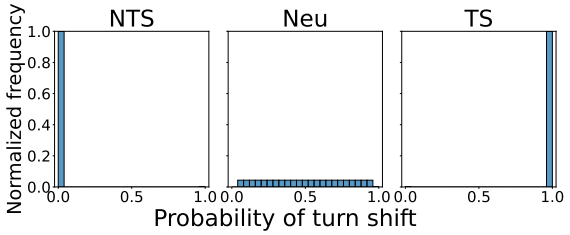


Figure 10: Uniform-shaped distributions

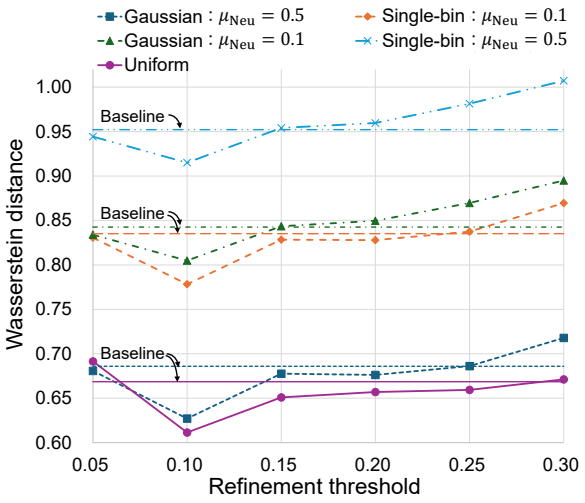


Figure 11: Performance across refinement thresholds under different reference distribution settings (FN Removal). Each horizontal line represents baseline performance corresponding to each reference distribution.

A Robustness to the Choice of Reference Distribution

We examined the robustness of the evaluation results under different reference distribution settings. For this analysis, FN Removal was used as a representative refinement strategy. We also examined how the relative performance of FN Removal against the baseline varied with the refinement threshold under these settings.

A.1 Settings of Reference Distributions

We selected five reference distribution settings to represent different assumptions about label ambiguity, as the shape of the reference distribution determines the penalties (or weights) used to measure distributional divergence. While it is intuitively reasonable to assume that the peaks (means) of the distributions for the *Non-Turn-Shift* (NTS) and *Turn-Shift* (TS) labels should lie near 0 and 1, respectively, the appropriate setting for the *Neutral* (Neu) label remains unclear. Therefore, we evaluated several representative parameter settings for the reference distributions.

As a baseline that provides smoothly varying (soft) penalties, we employed discrete Gaussian-shaped distributions with different means for the Neutral case, $\mu_{\text{Neu}} = 0.10$ and 0.50 (Figures 7 and 3). The value 0.10 was adopted as one candidate setting because it corresponds to the threshold selected in preliminary binary evaluation. By contrast, 0.50 represents a symmetric setting between the two binary endpoints. The standard deviations for the NTS, Neu, and TS labels were set to $\sigma_{\text{NTS}} = 0.027$, $\sigma_{\text{Neu}} = 0.100$, and $\sigma_{\text{TS}} = 0.093$, respectively.

We also introduced single-bin reference distributions (Figures 8 and 9), which impose hard penalties. These were derived from the Gaussian-shaped distributions (Figures 7 and 3) by taking $\sigma_{\text{NTS}}, \sigma_{\text{Neu}}, \sigma_{\text{TS}} \to 0$.

In addition, a uniform-shaped distribution (Fig-

ure 10) was adopted as an uninformative distribution for the Neutral label, assigning equal penalties across the probability range. This distribution can be regarded as a limiting case of the Gaussian-shaped distribution where $\sigma_{\text{NTS}}, \sigma_{\text{TS}} \to 0$ and $\sigma_{\text{Neu}} \to \infty$ over the range $(0.0, 1.0)$.

A.2 Results

Figure 11 shows the Wasserstein distance between the model outputs and the reference distributions across refinement thresholds under different reference distribution settings. The vertical axis represents the summed Wasserstein distance over all labels, while the horizontal axis represents the refinement threshold used for data refinement. We compared two alternative centers ($\mu_{\text{Neu}} = 0.1$ and 0.5) to represent a lower-probability assumption and a symmetric assumption about the Neutral label. Lower values indicate that the model output distribution is closer to the reference distribution. Solid lines represent FN Removal, while dashed lines represent the baseline. Lines with the same color correspond to the same reference distribution setting.

Across all reference distribution settings, the relative performance between the baseline and FN Removal was largely consistent. In all cases, the distance became smaller than the baseline around a threshold of 0.10, where it reached its minimum, and gradually increased as the threshold became larger, while remaining close to the baseline at other thresholds.

These results suggest that the main trend of the proposed evaluation was robust to the choice of reference distribution settings examined here. We further examined two design aspects of the reference distribution: its shape and the center of the Neutral label.

For the distribution shape, we compared Gaussian-shaped reference distributions with single-bin concentrated distributions. In both cases (centered at 0.5 and 0.1), the Gaussian-shaped distributions yielded smaller distances, suggesting that allowing a certain spread in the reference distribution was more compatible with the observed model outputs. While the uniform-shaped distribution showed broadly similar overall trends, it was treated as an uninformative control rather than a preferred target distribution for the Neutral label.

We also compared different centers for the Neutral label. Between Gaussian distributions centered at 0.5 and 0.1, the former yielded smaller distances. This suggested that a distribution centered between the peaks of *Non-Turn-Shift* and *Turn-Shift* was more appropriate.

Based on these observations, we used a Gaussian-shaped reference distribution with the Neutral distribution centered at 0.5 (Figure 3) in the main experiments.

B Additional Training Details

We experimented with different dialogue history lengths and found that performance stabilized when four preceding turns were included. Therefore, all experiments used a four-turn dialogue history.

The training configuration was as follows: the batch size was 2 with gradient accumulation over 10 steps (effective batch size: 20), the learning rate was 5×10^{-5}, and early stopping was applied when the validation loss did not improve for three consecutive epochs. We used the tokenizer provided with the model.

For overlap handling in preprocessing, we excluded instances in which more than half of the subsequent speaker's utterance overlapped with the preceding speaker, rather than treating them as turn-shift instances.

C Binary Evaluation Details

For binary evaluation, we used the subset of the annotated data for which a unique binary label could be defined. After excluding *Neutral* from the 500 annotated turns, 337 instances remained, of which 249 were labeled as *Turn-Shift* and 88 as *Non-Turn-Shift*.

We performed stratified two-fold cross-validation on these 337 instances, preserving the label distribution. In each fold, one split was used to select the optimal probability threshold for binary classification and the other for evaluation. The threshold was searched from 0.000 to 1.000 in increments of 0.001. Balanced accuracy (BAcc) was used because of the label imbalance.