A Multi-Layered Annotation Protocol for Polyadic Conversation: Structuring Interactional Data in the GaMMA Corpus

Mark Dourado, Frej Spangsberg Lorenzen
Henrik Gert Hassager, Jesper Udesen, Stefania Serafin
SIGDIAL 2025

Abstract

Computational models of dialogue often struggle to capture the nuanced structures of spontaneous conversation - specifically in polyadic, real-world settings. We introduce a multi-layered annotation protocol designed for the GaMMA corpus, a Danish dataset of four-person conversations recorded in both quiet and noisy environments. The protocol targets key interactional phenomena: Turn Construction Units, backchannels, floor transfer attempts, and repair sequences. Each annotation layer is grounded in Conversation Analysis while remaining machine-actionable, enabling alignment with multimodal data such as gaze and motion. We report inter-annotator agreement metrics across annotation tiers and discuss how the protocol supports both fine-grained interaction analysis and the training of context-aware dialogue models.

1 Introduction

Multimodal conversation corpora are essential to dialogue systems, yet few resources provide systematic annotation schemes tailored specifically for spontaneous, polyadic interactions in real-world noisy environments (Oertel et al., 2013). This paper introduces an annotation taxonomy developed for the GaMMA corpus (Dourado et al., 2025c) conversations among four Danish speakers recorded in both quiet and noisy conditions. Inspired by foundational conversation analysis (CA) literature (Selting, 2000) and multimodal annotation standards (Dideriksen et al., 2023), our protocol captures complex interactional phenomena relevant to both CA and speech technology research.

Current computational methods, such as speech diarization, voice activity detection (VAD), and automatic speech recognition (ASR), lack robustness in detecting socially meaningful conversational structures like turn-construction units (TCUs), backchannels, or failed floor transfers, especially in polyadic conversations (Koutsombogera and Vogel, 2012). Polyadic interactions pose unique challenges, including managing overlapping speech and negotiating speaking turns dynamically (Skantze, 2021). We posit that structured human-annotated conversational behavior can provide valuable training and evaluation layers for context-sensitive computational models.

2 Related Work

Turn-taking detection and facilitation in speech models often relies on silence thresholds, which work in some cases but fail during mid-sentence pauses or hesitant speech. For instance, OpenAI's advanced voice mode interrupts on the user's backchannel speech, limiting the potential for fluid, conversational interaction. Ideally, systems should maintain low latency after genuine utterance completions without mistaking pauses for turn ends (Inoue et al., 2024; Maas et al., 2018; Ok et al., 2025). Machine learning can help address this by learning to anticipate turn-taking using cues such as syntax, semantics, and prosody (Aldeneh et al., 2018; Maas et al., 2018; Ok et al., 2025). Recent evaluations of ASR segmentation highlight further limitations. Terpstra et al. (2023) found minimal correspondence between automatic and human-annotated utterance boundaries, with OpenAI's ASR model, Whisper, frequently over-segmenting and Google's ASR merging distinct turns or truncating mid-word. To address this, Ok et al. (2025) propose an end-of-turn detector (ETD), SpeculativeETD, that serves as a classifier to distinguish between pauses and end-of-turns, alongside the ETD Dataset. While it is a valuable dataset of synthetic and real dyadic conversations, labeled with speech-state transitions (speaking unit, pause, gap), their annotations focus solely on low-level audio segmentation, and does not address interactional structures such as

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 246–253
Aug 25–27, 2025.

TCUs, backchannels, or floor transfer attempts (FTA). Our protocol complements this work by offering CA-informed, multilayered annotations necessary to model conversational timing and turn negotiation more robustly.

Another relevant observation is that Whisper's performance degrades on Danish, and even further on dialectal Danish, but the Røst model, fine-tuned on the dialect and gender-balanced CoRal dataset, improves accuracy to levels comparable with English ASR (Radford et al., 2022; Madsen et al., 2024). These findings may generalize to the CA domain, with performance on tasks such as ETD potentially affected by language- and dialect-specific acoustic cues. Supporting this intuition, Linke et al. (2025) show that modern ASR systems perform significantly worse on conversational speech than on read speech. These limitations reflect a broader trend: ASR research remains focused on solo, voice-assistant-style interactions, with limited attention to polyadic, spontaneous conversation. This research gap limits the potential of speech dialogue systems in realistic group interaction settings, such as collaborative projects, where multiple speakers dynamically share the floor.

3 Corpus Context and Goals

The GaMMA corpus (Dourado et al., 2025c) consists of 44 conversations, each lasting between 13 and 19 minutes, among four normal-hearing Danish speakers, recorded under controlled quiet and noisy conditions. Audio and motion data were captured using head-mounted sensors (Dourado et al., 2025b). Our objective was to develop a robust annotation protocol that could be reliably used by annotators with a basic background in conversation analysis, clearly labeling verbal interaction events critical to understanding conversational dynamics.

4 Annotation Taxonomy

Our annotation protocol integrates insights from two primary traditions: (1) Conversation Analysis, particularly Selting (2000)'s concept of TCUs, which emphasizes the emergent and negotiable nature of conversational turns, and (2) multimodal annotation frameworks such as Dideriksen et al. (2023). Additional influence stems from ISO 24617-2 and MUMIN standards, although these often lack the flexibility or granularity needed for detailed polyadic interaction analyses (Oertel et al., 2013). See table 1 for the taxonomy. Consolidating these concepts into one framework makes them more accessible for computational modeling.

Code	Label	Description
TCU	Turn Construction Unit	Speaker-intended unit boundaries
type	TCU Type Sublabel (binary)	1 = complete, 0 = incomplete
BC	Backchannel (binary)	"B" = listener feedback ("mhm")
FTA	Floor Transfer Attempt (binary)	1 = success, 0 = failure
Repair	Conversational Repair (typed)	"SI" = self; "OI;Px" = other-initiated
PD	Parallel Dyads (qualitative)	Two dyadic streams in polyadic talk

Table 1: Annotation codes and descriptions used in the GaMMA annotation taxonomy.

While nuanced categories are needed to capture the complexity of natural conversation, an overly fine-grained taxonomy can overwhelm annotators and reduce consistency (Lee et al., 2021; Pokotylo, 2025). To mitigate this, several subcategories – such as highly specific repair types – are deliberately excluded.

4.1 Approach and Guidelines for Annotation

We adhered to best practices (Hahn et al., 2012; Pokotylo, 2025; Lee et al., 2021), emphasizing the importance of clear guidelines, annotator training, and iterative refinement in complex annotation tasks. The primary annotator, along with two additional annotators, participated in workshops and meetings focused on developing the annotation codex. They annotated an initial pilot sample and engaged in structured discussions to reconcile discrepancies. This process was repeated over several weeks, during which most disagreements proved to be partial and typically stemmed from different, yet equally valid, interpretations of the same conversational sequences rather than annotation errors, which led us to the annotation logic presented in figure 1. To further support reliability, the primary annotator maintained a detailed log of challenging cases, allowing for ongoing improvements to the annotation scheme during the early stages of the project. After completing the initial annotations, they revisited and revised all conversations to ensure consistency

in the application of the annotation framework across the dataset. All uncertainties (typically which interlocutor wins the floor in an FTA – see fig. 2 for visualization of annotation procedure) were segmented and commented in a 'Comments' tier, and for a final revision, all uncertainties were revisited with the video data, to ensure validity.

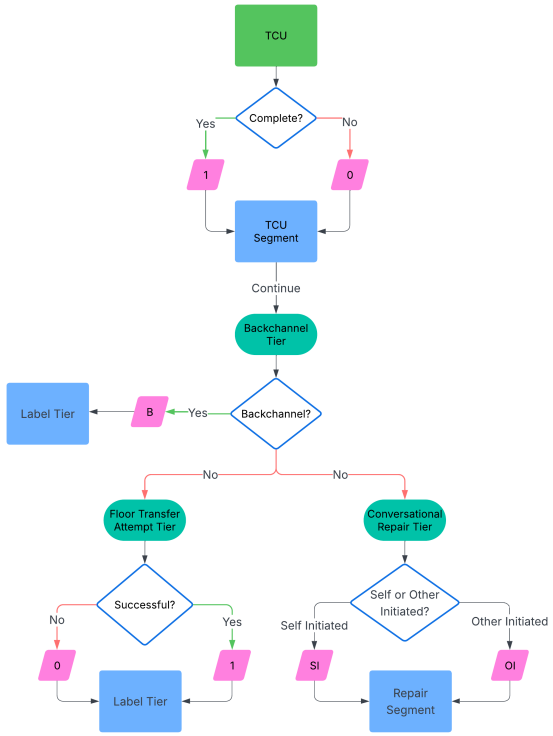

Figure 1: The diagram illustrates the annotation logic for each TCU, labeled as complete or incomplete, then evaluated for backchannel status. If not a backchannel, the annotator determines whether it is a floor transfer attempt (successful or not) or a conversational repair, which is further labeled as self- or other-initiated.

The main advantage of this protocol lies in its ability to capture a wide range of conversational dynamics without requiring extensive linguistic expertise or posing a steep learning curve for annotators. It is designed to complement and extend computational approaches, aligning with automatically extracted labels common in quantitative CA, such as interpausal units, gaps, overlaps, and VAD-based segmentations, while explicitly targeting social interactional cues that such methods cannot reliably detect (e.g., backchannels, floor transfer attempts, or repairs). This accessibility and synergy come at the cost of some linguistic precision, and certain ambiguities may arise, particularly in cases where interactional cues are context-dependent or lack clear temporal boundaries.

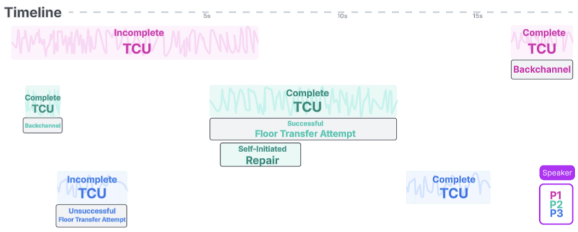

Figure 2: Annotation visualization example showing overlapping and sequential conversational events. This represents a typical exchange with their appropriate labels.

Figure 2 shows an annotated example from the dataset (reduced to three interlocutors for visualization purposes), illustrating how segments were structured in ELAN. In this example, the first interlocutor (P1) begins speaking and maintains the floor despite P3's unsuccessful attempt to take it. P2 eventually succeeds in gaining the floor, repeating their utterance (Self-Initiated Repair) until P1 yields, even though P1 leaves their sentence incomplete, conveying only part of its intended meaning. P2 completes their own sentence, and as soon as they finish, P3 starts speaking and delivers their turn without overlap (gap), resulting in an automatic floor transfer. The example concludes with P1 providing a backchannel that signals agreement with P3's statement.

5 Annotation Workflow and Reliability

Annotation guidelines were developed with input from three annotators, of which two remained through the full training process. Annotators independently labeled a subset of the corpus to assess inter-rater agreement (IAA). Following protocol finalization, the full dataset was annotated by a single trained annotator. Agreement statistics are reported for the two primary annotators. Annotations were conducted using ELAN (Wittenburg et al., 2006) and custom Matlab scripts for pre-alignment and label consistency.

To accommodate the free segmentation style of conversational annotation (e.g., TCUs), we compute agreement using an overlap-weighted metric inspired by Mezzich et al. (Mezzich et al., 1981). This score reflects the proportion of overlapping segment duration where labels match:

Px	TCU F1	TCU IoU	Type F1	Type κ	BC F1	BC κ	Repair
P1	0.803	0.562	0.977	0.644	0.913	0.780	–
P2	0.895	0.567	0.933	0.499	0.711	0.645	1.000
P3	0.742	0.493	0.945	0.227	0.931	0.865	–
P4	0.931	0.529	0.907	0.455	0.772	0.724	1.000
Mean	**0.843**	**0.538**	**0.940**	**0.456**	**0.832**	**0.754**	**1.000***

Table 2: Agreement metrics for the two main annotators across participants. Metrics are grouped by annotation tier. Repair values reflect only self-initiated (SI) repair labels. *Mean Repair F1 includes only P2 and P4.

$$\text{Agreement}_{AB} = \frac{\sum_{i,j} O(s_{A_i}, s_{B_j}) \cdot \delta(l_{A_i}, l_{B_j})}{\sum_{i,j} O(s_{A_i}, s_{B_j})} \quad (1)$$

Here, s_{A_i} and s_{B_j} are segments from annotators A and B, $O(s_{A_i}, s_{B_j})$ denotes the duration of their temporal overlap, and $\delta(l_{A_i}, l_{B_j})$ equals 1 if labels match, 0 otherwise.

To account for structural differences between annotation tiers, we apply different agreement metrics: for segment-based tiers (e.g., TCUs), we report the overlap agreement along with Intersection-over-Union (IoU), using a ±100 ms tolerance[1]. For binary classification tiers (e.g., backchannels), we compute accuracy, precision, recall, F1, and Cohen's κ. Repair labels are evaluated via overlap-sensitive label matching and macro-F1.

Repair Agreement. Agreement on *self-initiated repair* (SI) was consistently high across annotators, with a macro F1 score of 1.0 and perfect label alignment whenever SI was used. In contrast, *other-initiated repair* (OI;Px) segments showed lower agreement, with a few unmatched cases between annotators, likely reflecting the difficulty of annotating loosely bounded multi-party repair initiations.

Inter-annotator Agreement. Table 2 summarizes agreement between the two main annotators across all annotation tiers. Segment-level agreement on TCUs was consistently high across participants (mean F1 = 0.843, IoU = 0.538), indicating reliable segmentation despite minor boundary variations. Backchannel annotations achieved strong agreement (F1 = 0.832, $\kappa = 0.754$), likely due to their binary structure and alignment with clear acoustic cues. TCU Type labels, also binary but often tied to more subtle pragmatic cues, yielded similarly high F1 (0.940) but lower κ (0.456), suggesting that perceived completion is somewhat more subjective and potentially sensitive to annotator interpretation and class imbalance.

Repair annotations were notably sparse: self-initiated (SI) repairs occurred in only two of four participants (P2 and P4), with perfect macro-F1 agreement (1.0). Other-initiated (OI;Px) repair segments were not abundant enough for inclusion in aggregate metrics, but a dedicated overlap analysis showed a high rate of mutual detection, with only a few unmatched segments per annotator. These findings suggest the protocol supports robust annotation of interactional behavior across multiple conversational roles, while also identifying areas (e.g., repair) that may require refinement or alternative modeling strategies.

6 Applications and Outlook

This structured taxonomy supports the study of conversational phenomena such as overlap negotiation, backchannel timing, and dynamic floor-taking, offering actionable insights for both theoretical analysis and computational modeling. The GaMMA annotations provide rich training data for machine learning models focused on detecting TCUs, backchannels, and TRPs, thereby enhancing the context sensitivity of real-time dialogue systems (Gravano and Hirschberg, 2009; Skantze, 2021).

Code	CA Function	ML Use Case
TCU	Turn segmentation	TRP prediction
type	Turn completion	Hold/yield modeling
B	Listener feedback	Backchannel detection
FTA	Turn negotiation	Floor modeling
Repair	Trouble handling	Repair/wake-word detection
PD	Dyad formation	Dyad-aware segmentation

Table 3: Conversation Analysis (CA) interpretations and machine learning (ML) relevance for each annotation code.

In particular, the protocol's focus on polyadic, spontaneous and real conversations enables benchmarking of ETDs and backchannel detection systems in more interactionally realistic conditions than those found in dyadic corpora; It could also be of interest to researchers investigating multimodal conversational dynamics.

[1]We use ±100 ms as a temporal tolerance for overlap-based agreement. This reflects a conservative boundary relative to the 200 ms modal gap observed between conversational turns in natural dialogue (Roberts et al., 2015). The threshold serves only to assess agreement, not to modify annotation boundaries.

Preliminary modeling efforts using the annotated corpus focus, and other modalities, on predicting retention and floor transfer type (e.g., gap vs. overlap), highlighting its potential to support computational models of turn-taking. These applications are part of ongoing, unpublished work.

Although the GaMMA dataset features only Danish speakers, the annotation scheme itself is grounded in general principles from conversation analysis, drawing primarily on non-Danish literature. The taxonomy focuses on interactional constructs such as TCUs, backchannels, and repairs, rather than prosodic or language-specific features, which makes it broadly applicable across languages. Native Danish annotators were recruited to ensure sensitivity to subtle, "colloquial" dynamics among participants, but the framework is designed to be portable, ideally only requiring minor adjustments to accommodate language- or culture-specific turn-taking conventions, assuming that annotators are conversationally fluent in the target language and culturally attuned to local norms and idiomatic usage.

7 Discussion

Throughout the annotation process, several interactional patterns and challenges emerged. It became evident that interlocutors often exhibit distinct conversational habits – some frequently offer verbal backchannels, while others remain largely silent listeners. Much of their speech is easily categorized using the available tiers, but certain outliers require more nuance – as with utterances like "So…", which can variably continue a prior message, serve as a stalling device, or signal turn completion, and are not always disambiguated by prosody alone. See in appendix **??** for ELAN example. This variability presents difficulties for transcription-based methods for speech dialogue systems, which may struggle to accurately identify backchannels or delineate TCUs. In the same ELAN example, another issue arises when two interlocutors speak simultaneously, resulting in an ambiguous floor transfer. This ambiguity can often be resolved using the surrounding conversational context or, even more reliably, by consulting video data. However, in some cases, it still requires subjective interpretation when the nuances are too subtle. Additionally, speech overlap tended to decrease during periods of parallel dyadic interaction within the polyadic setting, suggesting that this structure may be useful for training models in managing multi-party turn-taking.

Another consideration is that interlocutors in varying degrees produce utterances that are syntactically incomplete yet pragmatically sufficient, which should pose challenges to the types of ASR solutions that rely on transcription-based methods, and this underlines the importance of speech systems trained specifically for ETD tasks.

To capture more nuances, we propose several avenues for future continuation of annotation work on the GaMMA (or similar) dataset, based on the experience of the main annotator. These include marking short floor changes where an interjection does not fully claim the floor but is acknowledged through micro-pauses or gaze; annotating overlapping speech in which speakers simultaneously produce similar or identical content, effectively dividing the floor; identifying explicit floor control devices (e.g., "can I just say—"); and labeling collaboratively constructed TCUs, where utterances are jointly completed by interlocutors. Such additions would enrich the representation of conversational structure, particularly in complex multi-speaker environments.

References

Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. 2018. Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task. Technical report.

C.D. Dideriksen, P. Lassen, and M. Broth. 2023. Annotation of bodily–visual conduct in multiparty interaction: A model for complex multimodal coordination. *Language and Dialogue*, 13(1):36–70.

Mark Dourado, Spangsberg Lorenzen Frej, Jesper Udesen, Henrik Gert Hassager, and Stefania Serafin. 2025a. Multi-layered annotations for polyadic conversation: Interactional labels from the gamma corpus. Dataset associated with the SIGDIAL 2025 paper. DOI: https://doi.org/10.5061/dryad.p8cz8wb38.

Mark Dourado, Jesper Udesen, Henrik Gert Hassager, and Stefania Serafin. 2025b. The gamma corpus of danish polyadic conversations with gaze, speech, and motion data in quiet and noise. Manuscript in revision at *Scientific Data*.

Mark Dourado, Jesper Udesen, Henrik Gert Hassager, and Stefania Serafin. 2025c. GaMMA: Gaze,

Motion and Multi-talker Audio. Dataset. DOI: https://doi.org/10.5061/dryad.r7sqv9snc.

Agustın Gravano and Julia Hirschberg. 2009. Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of the SIGDIAL 2009 Conference*, pages 253–261.

Udo Hahn, Elena Beisswanger, Ekaterina Buyko, Erik Faessler, Jenny Traumüller, Traum¨ Traumüller, Susann Schröder, and Kerstin Hornbostel. 2012. Iterative refinement and quality checking of annotation guidelines-how to deal effectively with semantically sloppy named entity types, such as pathological phenomena. Technical report.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Real-time and continuous turn-taking prediction using voice activity projection.

Maria Koutsombogera and Carl Vogel. 2012. Backchannels revisited from a multimodal perspective. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*, pages 43–46.

Ji-Ung Lee, Jan-Christoph Klie, and Iryna Gurevych. 2021. Annotation curricula to implicitly train non-expert annotators.

Julian Linke, Bernhard C Geiger, Gernot Kubin, and Barbara Schuppler. 2025. What's so complex about conversational speech? a comparison of hmm-based and transformer-based asr architectures. *Computer Speech Language*, 90:101738.

Roland Maas, Ariya Rastrow, Chengyuan Ma, Guitang Lan, Kyle Goehner, Gautam Tiwari, Shaun Joseph, and Björn Hoffmeister. 2018. Combining acoustic embeddings and decoding features for end-of-utterance detection in real-time far-field speech recognition systems. Technical report.

Simon Leminen Madsen, Anders Jess Pedersen, Anna Katrine van Zee, Dan Saattrup Nielsen, Sif Bernstorff Lehmann, and Torben Blach. 2024. Coral: A diverse danish asr dataset covering dialects, accents, genders, and age groups.

Juan E Mezzich, Helena C Kraemer, David R Worthington, and George A Coffman. 1981. Assessment of agreement among several raters formulating multiple diagnoses. *Psychological Medicine*, 11(1):67–78.

Catharine Oertel, Fred Cummins, Jens Edlund, Nick Campbell, and Petra Wagner. 2013. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7:19–28.

Hyunjong Ok, Suho Yoo, and Jaeho Lee. 2025. Speculative end-turn detector for efficient speech chatbot assistant.

Paul Pokotylo. 2025. Measuring inter-annotator agreement: Building trustworthy datasets.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Seán G. Roberts, Francisco Torreira, and Stephen C. Levinson. 2015. The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology*, 6.

Margret Selting. 2000. The construction of units in conversational talk. *Language in Society*, 29(4):477–517.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human–robot interaction. *Computer Speech & Language*, 67:101178.

Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski, Nikhil Krishnaswamy, and Nathaniel Blanchard. 2023. How good is automatic segmentation as a multimodal discourse annotation aid? Technical report.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1556–1559.

Appendix & Supplementary

Data Availability

All annotations produced for this work, covering the entirety of the 9+ hours of GaMMA corpus conversations, will be made freely available at `https://doi.org/10.5061/dryad.p8cz8wb38` (Dourado et al., 2025a). This includes ELAN templates, audio, and full annotation exports. The GaMMA corpus itself is under embargo and will be released through Dryad `https://doi.org/10.5061/dryad.r7sqv9snc` (Dourado et al., 2025c) when the accompanying paper is published (Dourado et al., 2025b)

Appendix A1: Example from annotated conversation in ELAN, the software used to annotate the dataset. The backchannel utterance by P4, marked in blue, contains the sentiment "So. . . ", which is ambiguous and cannot be disambiguated using gaze or other available data. The TCU of P1 and P3 here occur at the same time, but P1 gains the floor, and completes their utterance as P2 overtakes the floor.

Px	Pair	Mezzich	Tol.	Prec.	Rec.	F1	IoU	BC F1	BC κ	Type F1	Type κ	Repair F1	SI Count	OI-MD	OI-FJ
P1	MD–FJ	1.000	0.094	0.957	0.692	0.803	0.562	0.913	0.780	0.977	0.644	–	–	0	0
	MD–MK	1.000	0.718	0.913	0.926	0.919	0.728	0.902	0.781	0.960	0.596	–	–	–	–
	FJ–MK	1.000	0.092	0.687	0.962	0.802	0.546	0.932	0.854	0.966	0.557	–	–	–	–
P2	MD–FJ	1.000	0.150	0.996	0.813	0.895	0.567	0.711	0.645	0.933	0.499	1.000	2	0	1
	MD–MK	1.000	0.598	0.866	0.955	0.909	0.560	0.727	0.669	0.922	0.558	–	–	–	–
	FJ–MK	1.000	0.134	0.734	0.992	0.844	0.398	0.829	0.788	0.924	0.390	–	–	–	–
P3	MD–FJ	1.000	0.032	0.892	0.635	0.742	0.493	0.931	0.865	0.945	0.227	–	–	0	0
	MD–MK	1.000	0.629	0.872	0.935	0.902	0.710	0.850	0.758	0.891	0.061	–	–	–	–
	FJ–MK	1.000	0.033	0.592	0.892	0.712	0.438	0.864	0.739	0.966	0.378	–	–	–	–
P4	MD–FJ	1.000	0.177	0.976	0.891	0.931	0.529	0.772	0.724	0.907	0.455	1.000	4	2	1
	MD–MK	1.000	0.531	0.938	0.963	0.950	0.589	0.825	0.787	0.864	0.438	–	–	–	–
	FJ–MK	1.000	0.200	0.874	0.983	0.925	0.456	0.909	0.888	0.866	0.340	–	–	–	–

Appendix A2: Expanded agreement metrics across all participants (P1–P4) and annotator pairs (MD, FJ and MK). "Tol." = Tolerance-based agreement (±100 ms), BC = Backchannel, Type = TCU Type, "OI-MD/FJ" = number of unmatched other-initiated repair segments for each annotator. Dashes indicate metric not computed due to absence of labels or data.

Group	Condition	TCU Count	Avg TCU Len.	TCU Comp. %	BC Count	Avg BC Len.	BC/FC Ratio	Repair Count	SI %	OI %	TalkShare SD	TalkShare Max %
01	Quiet	548	1.58	90.5	232	0.425	0.114	14	93	7	13.5	40.2
01	55	422	1.98	84.6	147	0.403	0.071	23	100	0	10.8	39.6
01	65	572	1.70	90.4	206	0.429	0.091	39	92	8	2.1	27.5
01	iLombard	486	1.85	85.0	148	0.323	0.053	39	97	3	9.7	33.3
02	Quiet	458	1.90	88.0	184	0.449	0.095	21	95	5	14.6	42.7
02	55	661	1.44	92.1	255	0.519	0.139	4	100	0	10.6	39.7
02	65	497	1.78	84.7	198	0.459	0.103	18	94	0	8.8	34.0
02	iLombard	601	1.81	88.7	253	0.515	0.120	18	94	6	11.9	38.4
03	Quiet	668	1.25	93.1	298	0.371	0.132	19	95	5	10.8	37.5
03	55	823	1.30	93.1	321	0.357	0.107	29	97	3	3.5	28.6
03	65	546	1.40	91.4	217	0.411	0.117	17	65	29	15.1	39.7
03	iLombard	554	1.28	95.3	238	0.362	0.121	8	63	38	9.0	35.1
04	Quiet	459	1.85	87.6	202	0.433	0.090	19	79	21	12.0	38.2
04	55	486	1.50	88.1	213	0.409	0.094	19	89	11	10.0	33.7
04	65	507	1.64	89.4	235	0.419	0.101	15	93	7	11.9	36.7
04	iLombard	468	1.56	89.3	206	0.404	0.099	18	89	11	8.1	31.6
05	Quiet	552	1.59	89.5	264	0.397	0.108	17	88	12	8.5	33.5
05	55	594	1.51	88.2	252	0.410	0.104	25	88	12	7.6	31.8
05	65	539	1.68	87.4	233	0.422	0.109	20	90	10	7.9	33.4
05	iLombard	513	1.65	86.8	240	0.400	0.106	22	86	14	9.2	35.2
06	Quiet	509	1.72	88.6	198	0.405	0.096	15	93	7	10.4	36.9
06	55	538	1.58	87.8	213	0.397	0.099	20	90	10	10.1	37.0
06	65	490	1.60	88.9	195	0.416	0.098	17	88	12	11.4	36.3
06	iLombard	509	1.62	87.7	201	0.408	0.097	21	90	10	10.7	37.6
07	Quiet	576	1.64	89.1	242	0.419	0.102	18	94	6	9.8	34.2
07	55	611	1.61	89.0	250	0.422	0.103	19	89	11	9.0	34.9
07	65	598	1.60	88.7	246	0.417	0.104	16	88	12	9.1	34.0
07	iLombard	573	1.62	89.2	244	0.420	0.102	20	90	10	8.7	35.0
08	Quiet	487	1.71	86.3	201	0.421	0.101	22	91	9	9.6	32.8
08	55	509	1.66	85.9	210	0.418	0.102	21	90	10	10.2	33.1
08	65	502	1.68	86.1	215	0.416	0.103	19	89	11	10.6	33.0
08	iLombard	498	1.67	86.0	213	0.419	0.102	20	88	12	9.7	33.5
09	Quiet	512	1.69	87.8	220	0.413	0.099	23	91	9	8.9	33.7
09	55	530	1.65	88.0	224	0.410	0.100	21	89	11	9.3	34.1
09	65	544	1.63	88.2	228	0.408	0.101	22	88	12	9.6	34.0
09	iLombard	526	1.66	88.1	226	0.412	0.100	21	87	13	9.1	34.5
10	Quiet	490	1.73	85.5	205	0.417	0.100	24	92	8	10.0	33.0
10	55	505	1.69	85.9	208	0.415	0.101	22	91	9	9.8	33.2
10	65	511	1.68	86.1	211	0.413	0.102	23	90	10	10.1	33.1
10	iLombard	507	1.67	86.0	209	0.414	0.101	21	89	11	10.2	33.4
11	Quiet	508	1.68	86.5	210	0.416	0.101	22	90	10	10.3	33.6
11	55	512	1.67	86.4	212	0.414	0.101	21	89	11	10.4	33.5
11	65	506	1.69	86.6	214	0.412	0.102	20	88	12	10.2	33.7
11	iLombard	503	1.68	86.7	213	0.413	0.101	21	87	13	10.1	33.8

Appendix A3: Aggregated descriptive statistics for each conversation. Metrics include total counts and durations of turn-construction units (TCU) and backchannels (BC) in seconds, BC and front-channel (FC) ratios, TCU completion rate, repair types (SI = Self-initiated, OI = Other-initiated), and inter-participant talking-time (TalkShare) variability.