PAIR: A Pilot Dataset for Dual Perspective-based Video-Grounded Dialogue and Reconciliation

Lewis Watson, Carl Strathearn, Kenny Mitchell, Yanchao Yu
Edinburgh Napier University
10 Colinton Road, Edinburgh, EH10 5DT
{L.Watson, K.Mitchell2, C.Strathearn, Y.Yu}@napier.ac.uk

Abstract

Collaborative dialogue in multi-agent settings often requires interlocutors to integrate partially overlapping perceptual information in order to construct a shared representation of a dynamic environment. We introduce PAIR, a pilot conversational corpus designed to examine how humans coordinate under systematic perceptual asymmetry. The dataset comprises 15 dialogues in which participants observed the same activity from complementary egocentric and exocentric video perspectives and engaged in open-ended discussion to produce a joint account. All transcripts were manually verified and annotated with 42 dialogue act categories, enabling fine-grained analysis of interactional structure. Beyond descriptive statistics, PAIR supports examination of measurable conversational configurations, including turn distribution, participation symmetry, and dialogue act composition, which together provide structural indicators of how perspective integration unfolds in dialogue. Although intentionally lightweight, PAIR is positioned as a controlled benchmark for analysing collaborative dialogue mechanisms rather than a large-scale training resource. The corpus supports dialogue act classification, video-grounded dialogue modelling, and investigation of multi-agent reasoning under distributed perceptual access. By coupling dual-perspective grounding with explicit interactional annotation, PAIR offers a compact testbed for studying reconciliation dynamics in task-oriented dialogue.

Keywords: dual-perspective dialogue, video-grounded dialogue, perceptual asymmetry, participation symmetry, dialogue act annotation, multi-agent collaboration

1. Introduction

Collaborative dialogue frequently requires participants to integrate partially overlapping perspectives to construct a shared representation of a dynamic environment. In many real-world settings—such as human–robot interaction, multi-agent coordination, or remote collaboration—interlocutors do not have identical perceptual access to the scene under discussion. Instead, they must reconcile complementary and sometimes conflicting observations through iterative clarification, questioning, and reformulation. Despite the importance of such distributed perception, most existing dialogue corpora either assume shared visual grounding or adopt rigid question–answer formats that limit analysis of open-ended perspective integration.

PAIR is a pilot conversational corpus designed to examine how humans negotiate shared understanding in the presence of perceptual asymmetry. The dataset comprises 15 dialogues in which participant pairs discuss the same event from distinct viewpoints: one observes an egocentric (first-person) recording, while the other views a corresponding exocentric (third-person) recording of the same activity. This dual-perspective design systematically induces complementary access to visual information. All conversations were recorded, transcribed using WhisperX (Bain et al., 2023), manually verified for accuracy, and annotated with 42 dialogue act categories to capture interactional structure at fine granularity.

Speaker 1	Speaker 2
	

S_0: I wasn't actually sure what it was. *(STATEMENT)*
S_1: And coriander at the end? *(YES-NO-QUESTION)*
S_0: Coriander at the end. *(REPEAT-PHRASE)*
S_1: Is that the last thing? *(YES-NO-QUESTION)*
S_0: Yeah. *(YES ANSWERS)*
S_1: It was like a garnish? *(DECLARATIVE Y-N-QUESTION)*
S_0: That's right. *(AGREEMENT/ACCEPT)*

Figure 1: Example dialogue from the cooking video clip in the Ego-Exo4D dataset (Grauman et al., 2024)

Unlike image-based Q&A datasets or visually grounded dialogue corpora that constrain conversational flow, PAIR enables analysis of perspective integration as an interactional process. In addition to providing transcripts and dialogue act annotations, the dataset supports structural examination of turn distribution, participation symmetry, and dialogue act sequencing. These measurable features allow reconciliation to be operationalised in terms of observable conversational configurations rather than inferred solely from task outcomes.

PAIR is intentionally lightweight and positioned as a controlled benchmark rather than a large-scale training resource. Its compact scale enables close analysis of interactional mechanisms, including how floor exchange patterns align with questioning behaviour and how participation symmetry relates to clarification dynamics. The corpus supports tasks such as dialogue act classification, video-grounded dialogue modelling, and structural analysis of collaborative scene reconstruction.

PAIR[1] is released to facilitate research on perspective-based dialogue and multi-agent reasoning. As a proof-of-concept resource, it provides both an empirical foundation for studying reconciliation dynamics and a methodological template for extending dual-perspective dialogue collection to broader multimodal and embodied interaction settings.

2. Related Work

This section critically reviews corpora that model collaborative human dialogue and situates PAIR within this landscape. We compare prior work along seven structural dimensions: (1) dynamic visual grounding, (2) dual or asymmetric perceptual access, (3) free-form multi-turn dialogue, (4) task-oriented collaboration, (5) explicit reconciliation under uncertainty, (6) dialogue act annotation, and (7) multimodal reconstructability (see Table 1).

Early work produced large-scale conversational datasets without perceptual grounding. The Switchboard Corpus (Godfrey et al., 1992) captures telephone conversations annotated with dialogue acts and remains foundational for modelling turn-taking and conversational structure. However, it lacks environmental grounding, task collaboration, and perceptual asymmetry. Dialogue is interactional but not jointly problem-solving in a shared scene.

Task-oriented corpora embed dialogue within structured problem-solving contexts. The Map-Task Corpus (Anderson et al., 1991) introduces partial information asymmetry, requiring speakers to align spatial maps, while the TRAINS corpus (Allen et al., 1995) records collaborative planning dialogues in a logistics domain. The AMI Meeting Corpus (Carletta et al., 2005) extends collaboration to multi-party settings with multimodal recordings, enabling analysis of group decision-making dynamics. Negotiation and game-based datasets (Guhe and Lascarides, 2014; Lewis et al., 2017) further examine strategic interaction and persuasion. These resources foreground task collaboration and interactional structure, yet they typically assume either symbolic tasks or shared perceptual access rather than dynamic visual asymmetry.

[1]Available in repository https://github.com/lewiswatson55/PAIR_Corpus

Other corpora ground dialogue in interactive environments. The GIVE Challenge (Gargett et al., 2010) links instruction-following language to virtual world actions, and TaskMaster (Strathearn et al., 2023) explores context-rich spoken dialogue for task completion. While grounded in action and problem-solving, these datasets rely on shared environmental context or single-viewpoint perception.

Vision-language datasets integrate perceptual input directly. Visual Dialog (Das et al., 2017) and GuessWhat?! (De Vries et al., 2017) connects dialogue to images, enabling multi-turn reasoning over visual content. However, interactions are typically structured as question–answer exchanges rather than collaborative negotiation. Similarly, the PhotoBook dataset (Haber et al., 2019) investigates visually grounded reference dialogue and the incremental establishment of common ground. Although PhotoBook introduces partial perceptual differences between participants, it relies on static images and does not incorporate dynamic video-based asymmetry.

Across these corpora, perceptual grounding, task collaboration, and dialogue structure are often studied in isolation. Most assume shared perceptual access to the same scene, even when dialogue is collaborative.

Unlike existing corpora that assume shared perceptual access, PAIR systematically induces perceptual asymmetry, requiring participants to resolve potentially conflicting observations. Participants observe the same dynamic event from egocentric and exocentric viewpoints and must reconcile incomplete or overlapping information into a coherent shared account. This design simultaneously integrates dynamic video grounding, dual-perspective perception, free-form dialogue, explicit reconciliation, and fine-grained dialogue act annotation.

Though lightweight, PAIR occupies a distinct position within the dialogue dataset landscape: it combines perceptual asymmetry with collaborative problem-solving under dynamic visual conditions. As such, it provides a controlled benchmark for modelling perspective reconciliation and multi-agent reasoning grounded in real-world video contexts.

3. PAIR Knowledge-Sharing Task

Task & Materials The PAIR dataset is constructed around a controlled knowledge-sharing task designed to elicit perspective reconciliation under asymmetric perceptual access. Video stimuli were drawn from the Ego–Exo4D dataset (Grauman et al., 2024), which records skilled human activities simultaneously from egocentric (first-person) and exocentric (third-person) viewpoints. This dual-view property enables controlled induction of per-

Dataset	Dynamic Video	Dual Perspective	Free Dialogue	Task-Oriented	Explicit Reconciliation	Dialogue Acts	Multimodal Reconstructable
Switchboard (Godfrey et al., 1992)	–	–	✓	–	–	✓	–
MapTask (Anderson et al., 1991)	–	Partial	✓	✓	Partial	✓	–
AMI (Carletta et al., 2005)	–	–	✓	✓	Partial	✓	✓
Visual Dialog (Das et al., 2017)	Image	–	QA-style	–	–	–	✓
GuessWhat?! (De Vries et al., 2017)	Image	–	Structured QA	✓	Limited	–	✓
CLEVR-Dialog (Kottur et al., 2019)	Image	–	Structured	✓	–	✓	✓
VDAct (Imrattanatrai et al., 2025)	✓	–	✓	✓	Limited	–	✓
TikTalk (Lin et al., 2023)	✓	–	✓	–	–	–	✓
PhotoBook (Haber et al., 2019)	Image	✓	✓	✓	✓	–	✓
HowToDIV (Aggarwal et al., 2025)	✓	–	✓	✓	Limited	–	✓
PAIR (Ours)	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of PAIR with representative dialogue corpora across seven structural dimensions. PAIR uniquely combines dynamic dual-perspective video grounding, free-form collaborative dialogue, explicit reconciliation, and fine-grained dialogue act annotation.

ceptual asymmetry: while both participants observe the same underlying event, each has access to distinct and partially overlapping visual evidence.

Five activity clips (approximately five minutes each) were selected to represent diverse task types (e.g., cooking, sports, cycling). Although the selection is randomised, we tried to keep balance between different tasks in the experiment.

In each session, two participants simultaneously viewed the same clip from different vantage points: one watched the egocentric recording, and the other, the exocentric recording. Participants were instructed not to communicate during viewing. Immediately afterwards, they engaged in an unconstrained face-to-face dialogue to produce a shared account of the observed activity. No structured script was provided beyond the instruction to "work together to agree on what happened in the video." This open-ended format was intended to elicit natural collaborative reasoning rather than constrained question–answer interaction.

Participants were seated opposite each other and permitted to use natural conversational resources (e.g., gesture, gaze, backchanneling). However, only the audio transcription was recorded and preserved in the released dataset. After participants indicated that they had reached an agreement, they were asked to produce a brief joint sketch summarising the activity. These sketches were used solely to encourage explicit consensus formation and were not retained for analysis.

This protocol ensures that reconciliation is structurally required: neither participant possesses a complete representation of the scene, and shared understanding must emerge through dialogue.

Participants Thirty participants were recruited from a university population. Participants were randomly paired to minimise familiarity effects. Each of the five video conditions was completed by three independent pairs, yielding 15 dialogues in total. All participants were fluent English speakers.

The repeated-pair design allows analysis of variation in conversational strategy under identical perceptual conditions, enabling comparison between interactional styles while holding visual input constant. All procedures were conducted under institutional ethical approval, and participants provided informed consent before data collection.

Transcription, Verification & Annotation Each collaborative discussion lasted approximately 10 minutes and was recorded using a centrally positioned audio recorder. Initial transcripts were generated using WhisperX (Bain et al., 2023), which integrates automatic speech recognition with speaker diarization built on Whisper (Radford et al., 2022).

Although word-level transcription accuracy was generally high, diarization errors were common. All transcripts were therefore manually reviewed and corrected, including:

- Verification of speaker labels,
- Correction of misrecognised tokens,
- Removal of non-linguistic artefacts where appropriate.

Utterances were segmented using a 500 ms pause threshold to ensure consistent turn boundaries across dialogues.

All utterances were annotated using the 42 dialogue act categories defined by Stolcke et al. (2000) (see annotation schema in Appendix A[2]). The schema captures both informational content (e.g., *STATEMENT*, *YES-NO-QUESTION*) and interactional mechanisms (e.g., *HEDGE*, *COLLABORATIVE COMPLETION*, *RESPONSE ACKNOWLEDGMENT*). This fine-grained annotation enables

[2]The annotation schema was discussed and refined collaboratively among the annotators and research team before corpus annotation to ensure consistent interpretation of dialogue act definitions.

analysis not only of descriptive content but also of the mechanisms through which participants negotiate uncertainty, repair misunderstandings, and establish alignment.

4. The Corpus Analysis

The **PAIR corpus** in this study is derived from audio recordings of participants who were tasked with discussing and interpreting a perceptual scene from their unique perspectives.

# dialogues	15
# dialogues per video clip	3
Total # of utterances	2676
Avg. # words per utterance	8.61
Total # of turns (no filler)	2027
Min. # turns per dialogue	77
Avg. # turns per dialogue	135.13
Median conversation length (sec)	636

Table 2: Properties of Dialogue Corpus

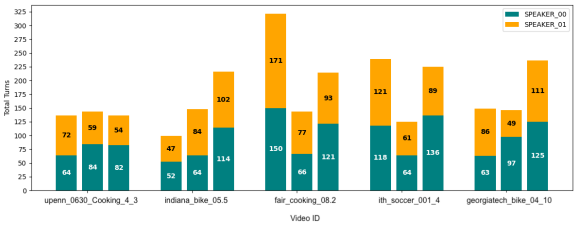
Table 2 summarises the key properties of the PAIR dataset. A total of 15 dialogues were collected across five video scenarios, with each scenario repeated by three different participant pairs. The corpus contains $2,676$ utterances and $2,027$ conversational turns, excluding filler speech. An *utterance* is defined as a continuous segment of speech separated by at least 500 milliseconds of silence, while a *turn* represents the transfer of the conversational floor between speakers.

On average, each dialogue contains around 178 utterances and 135 turns, with the shortest dialogue including 77 turns. Utterances average 8.61 words in length, indicating concise contributions that nonetheless build up into complex exchanges. The median conversation length is 636 seconds (approximately 10.6 minutes), reflecting the consistent design of the knowledge-sharing task.

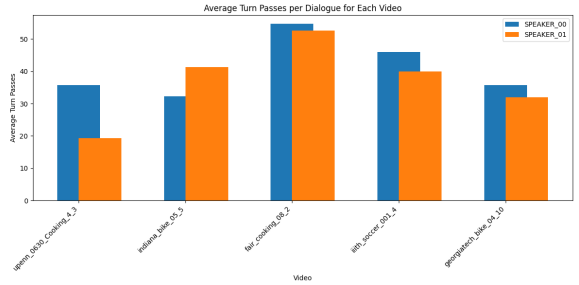
4.1. Dialogue Structuring: Utterance Length vs. Interaction Density

Figure 3 reports two descriptive statistics for each video scenario in the PAIR dataset: the *average utterance length* (blue line, left axis) and the *average number of utterances per dialogue* (green line, right axis). These measures characterise how conversational contributions are distributed within each task context.

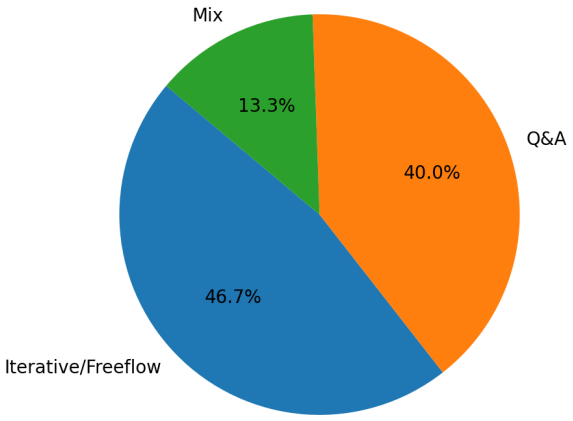
Across scenarios, an inverse relationship between utterance length and frequency is observed at the aggregate level. For example, in *upenn_0630_cooking_4_3*, dialogues contain the longest utterances (over 54 characters on average) but the fewest total utterances (approximately 137 per conversation). In contrast, *fair_cooking_08_2* exhibits shorter utterances (around 41 characters



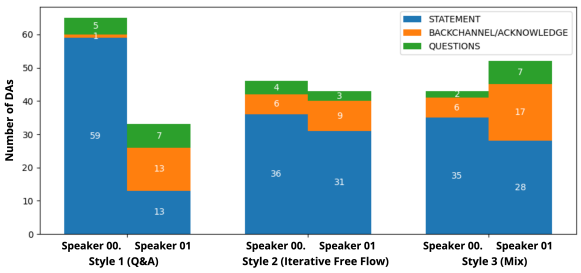
(a) Turn Distribution between two speakers per dialogue



(b) Turn Passes per Dialogue for each Speaker



(c) Interaction Styles Distribution Across Dialogues



(d) DA Distribution Across Interaction Styles

Figure 2: Corpus Statistics

on average) alongside the highest number of utterances (over 220 per dialogue). Other scenarios, such as *iiith_soccer_001_4* and *georgiatech_bike_04_10*, fall between these extremes, with moderate utterance lengths and mid-range utterance counts.

These descriptive patterns indicate that different task contexts are associated with distinct interactional distributions. Some dialogues feature fewer but longer contributions, whereas others feature

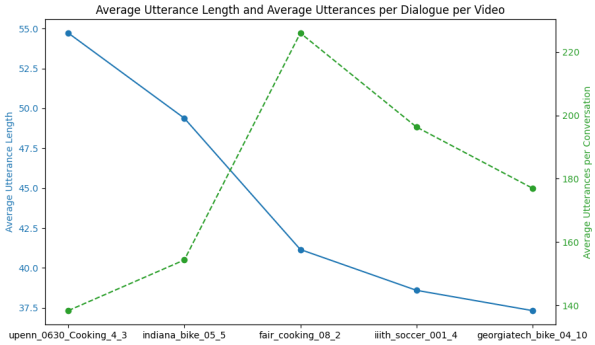


Figure 3: PAIR Dialogue Structuring

more frequent but shorter exchanges. However, given the limited number of dialogues per scenario (three instances each), these observations should be interpreted cautiously as corpus-level tendencies rather than statistically validated behavioural effects.

Importantly, utterance length and utterance frequency alone do not determine communicative strategy or collaborative efficiency. They provide structural indicators of conversational organisation but do not directly reveal whether exchanges reflect elaboration, uncertainty, clarification, or coordination. Further analysis incorporating dialogue-act distributions and sequential patterns would be required to substantiate stronger claims about adaptive communicative behaviour.

In this sense, Figure 3 should be understood as a structural characterisation of dialogue density across task types, rather than evidence of causal adaptation to perceptual or cognitive demands.

4.2. Interaction Styles Across Dialogues

To characterise broader conversational organisation, we examined interaction styles based on relative speaker contributions and dialogue-act distribution. Using proportional turn distribution and dominant dialogue act types, three recurrent configurations were identified across the 15 dialogues:

- **Q&A Style (40%):** Marked by asymmetric participation, where one speaker contributes a higher proportion of *STATEMENT* acts while the other produces more question and acknowledgement acts.
- **Iterative Freeflow Style (46.7%):** Characterised by relatively symmetric turn distribution and balanced proportions of statements and backchannels across speakers.
- **Mix Style (13.3%):** Displays moderate asymmetry, with one speaker contributing more extended statements while the other remains actively engaged through questioning and acknowledgement.

Figure 2c presents the distribution of these styles, and Figure 2d shows the corresponding dialogue act breakdown. Q&A and Freeflow together account for approximately 87% of dialogues, indicating that most sessions fall into either clearly asymmetric or relatively symmetric participation configurations.

These style categories describe structural interactional patterns rather than qualitative assessments of collaboration. Freeflow dialogues exhibit higher participation symmetry and mutual statement construction, whereas Q&A dialogues concentrate extended declarative contributions in one speaker. The Mix style falls between these configurations.

Compared with datasets that enforce rigid question–answer formats (e.g., VQA, Visual Dialog), PAIR contains a higher proportion of dialogues with bidirectional exchange and distributed questioning. This structural diversity supports analysis of perspective integration as an interactional process rather than a fixed prompt–response format[3].

4.3. Interaction Intensity Across Videos

Figure 2a reports the total number of conversational turns per video scenario, separated by speaker. A *turn* is defined as a transfer of the conversational floor. Turn counts therefore provide a structural indicator of interaction density and participation symmetry rather than a direct measure of collaboration quality.

To quantify participation balance, we compute a *Turn Balance Ratio* (TBR):

$$\mathrm{TBR} = \frac{\min(T_{S0}, T_{S1})}{\max(T_{S0}, T_{S1})}, \tag{1}$$

where T_{S0} and T_{S1} denote the number of turns produced by each speaker. $\mathrm{TBR} \in (0, 1]$, with values closer to 1 indicating more symmetric participation.

Interaction density varies across scenarios. The *fair_cooking_08_2* condition exhibits the highest overall activity (321 turns) with relatively balanced contributions (150 vs. 171; $\mathrm{TBR} = 0.88$). In contrast, *indiana_bike_05_5* shows fewer turns and greater asymmetry (102 vs. 64; $\mathrm{TBR} = 0.63$). The *iith_soccer_001_4* dialogues contain over 200 turns with near parity between speakers, while *georgiatech_bike_04_10* displays moderate totals with observable variation across sessions.

Across scenarios, both total turn counts and TBR values differ, indicating that interaction density and

[3]Given the pilot scale of the dataset (15 dialogues), these style distinctions should be interpreted as descriptive corpus patterns rather than statistically validated categories. Larger samples will be required to test their stability and their relationship to reconciliation dynamics formally.

participation symmetry are not uniform across task contexts. Cooking scenarios show comparatively high turn counts and higher balance ratios, whereas some biking scenarios display lower density and greater asymmetry. Sports-related dialogues exhibit moderate-to-high interaction density with relatively symmetric participation.

These measures characterise structural organisation of dialogue rather than underlying communicative intent. Turn frequency and balance do not, by themselves, distinguish coordination from uncertainty, disagreement, or redundancy. Interpretation of collaborative dynamics therefore requires integration with dialogue act sequencing and content-level analysis.

The present findings should be understood as corpus-level structural differences across task settings. Establishing whether task characteristics systematically influence participation patterns will require statistical validation on larger samples and integration with finer-grained interactional metrics.

4.4. Turn-Passing Patterns

Figure 2b reports the average number of turn passes per dialogue across five video scenarios, separated by speaker. A *turn pass* occurs when one speaker yields the conversational floor, and the other assumes it. Unlike utterance counts, which capture speech segments, turn passes characterise the frequency of floor exchange between participants and therefore provide a structural measure of interactional coordination.

Turn-passing frequency varies across scenarios. In *upenn_0630_cooking_4_3*, S0 produces substantially more turns than S1, indicating asymmetric floor control. In *indiana_bike_05_5*, the asymmetry is reversed, with S1 contributing more turns. The most balanced exchanges appear in *fair_cooking_08_2*, where both speakers produce comparable numbers of turns (over 50 each). *iiith_soccer_001_4* and *georgiatech_bike_04_10* also exhibit near-parity, although slight dominance by one speaker is observable in some sessions.

These differences indicate that floor exchange patterns are not uniform across task settings. Some dialogues are characterised by asymmetric turn-taking, while others display frequent, relatively balanced exchanges. However, turn-passing frequency alone does not determine collaboration quality. High exchange rates may reflect coordinated negotiation, iterative clarification, or uncertainty-driven repair, whereas asymmetric patterns may arise from narrative style, role adoption, or temporary dominance within the interaction. Accordingly, turn-passing should be interpreted as a structural property of dialogue organisation rather than a direct proxy for collaborative effectiveness.

Dialogues with more symmetric turn distributions tend to contain greater proportions of mutual statement and acknowledgement acts. In contrast, dialogues with more asymmetric turn distributions show a higher concentration of extended STATEMENT acts by one speaker, accompanied by a greater proportion of question acts from the other to extract information. These observations indicate that floor exchange patterns correspond to differences in interactional function at the level of dialogue acts.

To quantitatively validate the relationship observed in our style classification (see Section 4.2), we computed the Spearman correlation between TBR and the proportion of question-related dialogue acts per dialogue (grouping YES-NO-QUESTION, WH-QUESTION, DECLARATIVE YES-NO-QUESTION, OPEN-QUESTION, and related categories; see Appendix B). Across the 15 dialogues, TBR showed a significant strong negative association with question proportion ($\rho = -0.7120$, $p = 0.002905$). Although the sample is limited, this negative correlation indicates that a more symmetric floor distribution (higher TBR) is actually associated with a lower relative frequency of questioning behaviour. This aligns with our style classifications: highly symmetric dialogues consist of participants mutually trading statements and agreements to build a shared account, whereas asymmetric dialogues rely heavily on one participant driving the interaction through concentrated questioning.

Overall, turn-passing variation across scenarios reflects differences in conversational organisation. Establishing whether these differences systematically relate to task characteristics or perceptual asymmetry will require an integrated analysis combining floor-exchange metrics with sequential dialogue-act modelling.

4.5. Structural Indicators of Reconciliation Dynamics

Taken together, interaction density (total turns), participation symmetry (TBR), and dialogue act distributions provide complementary structural indicators of how shared understanding is negotiated.

Dialogues with higher TBR values exhibit frequent bidirectional exchange driven primarily by mutual statements and acknowledgements. In contrast, dialogues with lower TBR values tend to concentrate extended STATEMENT acts in one speaker, with the other participant relying on a greater proportion of question acts to extract information. While these measures do not directly capture successful reconciliation outcomes, their co-occurrence patterns suggest two primary structural strategies for reconciliation, alongside a hybrid approach. In this sense, reconciliation in PAIR

can be operationalised not as a binary outcome, but as varying interactional configurations, ranging from mutually constructed narratives (Iterative Freeflow Style) to asymmetrical, clarification-driven interviews (Q&A Style), with some dialogues exhibiting a mix of both dynamics.

This framing positions PAIR as a resource for analysing reconciliation mechanisms through measurable dialogue structure, rather than inferring collaboration solely from task completion.

5. Conclusion & Future Work

PAIR is a pilot conversational corpus developed to examine how humans integrate complementary perspectives when interpreting dynamic visual scenes. With fifteen dialogues, the dataset is intentionally lightweight and positioned as a controlled proof-of-concept resource. Rather than emphasising scale, PAIR prioritises structural depth through dual-perspective video grounding, fine-grained dialogue act annotation, and manually verified transcripts. Its primary contribution lies in providing a compact benchmark for analysing mechanisms of collaborative dialogue under perceptual asymmetry.

This work introduces a structural approach to examining reconciliation in dialogue. As demonstrated in Section 4, participation symmetry, turn-exchange patterns, and dialogue-act configurations together provide measurable indicators of how perspective integration unfolds. Rather than treating reconciliation as an abstract outcome, PAIR enables it to be analysed through observable interactional structure.

Future extensions should incorporate multimodal recordings, including gaze, gesture, and facial expression, to capture additional coordination signals. Increasing activity diversity and sample size will allow for a more robust statistical evaluation of the structural tendencies identified here. Overall, PAIR provides a foundation for benchmarking how human and artificial agents reconcile distributed viewpoints and coordinate toward shared representations in dynamic, task-oriented environments.

Ethics Statement

All participant recruitment and data collection for the PAIR corpus were conducted under institutional ethical approval. Before participation, all individuals provided written informed consent after being fully briefed on the study aims, recording procedures, data handling, and their right to withdraw without consequence.

All recordings were made specifically for research purposes and processed in accordance with GDPR and institutional data-protection policy. Only anonymised audio transcripts and derived annotations are shared publicly; no identifiable video, audio, or personal information is released.

The video stimuli used in this study were drawn from the *Ego-Exo4D* dataset (Grauman et al., 2024), which is publicly available under a research-only licence. In compliance with its terms, the original *Ego-Exo4D* video data are not redistributed as part of PAIR. Instead, PAIR includes only dialogue transcripts, dialogue-act annotations, and metadata derived from those sessions. Users must independently obtain access to the *Ego-Exo4D* videos through the official repository for any reproduction or re-analysis.

This work aims to advance transparent and responsible research in human-AI collaboration. The dataset is intended exclusively for academic use and should not be used for surveillance, profiling, or other purposes that may infringe on participants' privacy or autonomy. Researchers are encouraged to follow FAIR and ethical-use principles when extending or adapting PAIR.

Limitations

While PAIR offers a novel testbed for studying dual-perspective, video-grounded dialogue, it remains a pilot-scale corpus with several important limitations.

- **Scale and Behavioural Coverage:** PAIR comprises 15 dialogues across five activity types. Although each scenario was repeated three times to capture intra-task variation, the overall corpus size limits statistical robustness. It does not exhaust the full spectrum of collaborative strategies observable in natural dialogue. The dataset is therefore better suited to controlled analysis of reconciliation mechanisms than to large-scale model training or broad generalisation claims.

- **Demographic Diversity:** Participants were recruited primarily from a university-affiliated population. As a result, the corpus reflects a relatively narrow range of age groups, educational backgrounds, and cultural-linguistic profiles. Collaborative dialogue patterns may differ across broader populations, and future expansions should incorporate more diverse participant groups to improve representativeness.

- **Modal Scope:** Although the interactions were conducted face-to-face and included gesture, gaze, and other non-verbal cues, only audio recordings and manually verified transcripts are released. Consequently, analyses are limited to verbal behaviour, and multimodal coordination signals central to perspective reconciliation are not preserved in the current release.

- **Task Constrainedness:** The knowledge-sharing task was conducted under controlled laboratory conditions using selected *Ego-Exo4D* clips. While this ensures consistency and comparability across sessions, it also constrains the interactional space. Real-world collaborative settings may involve more open-ended goals, shifting social dynamics, and richer multimodal environments.

Taken together, these design choices position PAIR as a controlled, mechanistic benchmark for studying perspective reconciliation rather than a comprehensive coverage corpus. Its strength lies in deliberately inducing perceptual asymmetry under comparable conditions. Future extensions will increase dialogue volume, diversify participant demographics, and incorporate multimodal recordings to enable broader generalisation and more ecologically grounded evaluation of collaborative dialogue systems.

Acknowledgements

We thank the anonymous reviewers and the area chair for their constructive comments. This research is supported by the EPSRC, under grant number UKRI210 (LoCS project).

6. References

Lavisha Aggarwal, Vikas Bahirwani, Lin Li, and Andrea Colaco. 2025. Generating dialogues from egocentric instructional videos for task assistance: Dataset, method and benchmark. *arXiv preprint*.

James F Allen, Lenhart K Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel Martin, Bradford Miller, Massimo Poesio, et al. 1995. The trains project: A case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):7–48.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and Speech*, 34(4):351–366.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 28–39. Springer.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The GIVE-2 corpus of giving instructions in virtual environments. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE.

Kristen Grauman, Andrew Westbury, et al. 2024. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives.

Markus Guhe and Alex Lascarides. 2014. Game strategies for the settlers of catan. In *2014 IEEE Conference on Computational Intelligence and Games*, pages 1–8. IEEE.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The photobook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.

Wiradee Imrattanatrai, Masaki Asada, Kimihiro Hasegawa, Zhi-Qi Cheng, Ken Fukuda, and Teruko Mitamura. 2025. A video-grounded dialogue dataset and metric for event-driven activities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24203–24211.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in

visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 582–595.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

Hongpeng Lin, Ludan Ruan, Wenke Xia, Peiyu Liu, Jingyuan Wen, Yixin Xu, Di Hu, Ruihua Song, Wayne Xin Zhao, Qin Jin, et al. 2023. Tiktalk: A video-based dialogue dataset for multi-modal chitchat in real world. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1303–1313.

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

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Carl Strathearn, Yanchao Yu, and Dimitra Gkatzia. 2023. Taskmaster: A novel cross-platform task-based spoken dialogue system for human-robot interaction. In *Proceedings of the Workshop on Human-Robot Conversational Interaction (HRCI) at the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Stockholm, Sweden.

A. Tagging Schema

Table 3 lists the 42 dialogue act categories applied to PAIR, following the schema introduced by Stolcke et al. (2000). Each label corresponds to a distinct conversational function, ranging from core informational acts (e.g., *STATEMENT*, *YES-NO-QUESTION*) to interactional management acts (e.g., *BACKCHANNEL/ACKNOWLEDGE*, *HEDGE*, *COLLABORATIVE COMPLETION*).

For each category, an example utterance from PAIR is provided to illustrate its usage in context. Some dialogue acts defined in the original schema (e.g., *TAG-QUESTION*, *APOLOGY*) do not appear in the corpus and are indicated accordingly.

The use of this fine-grained tagging scheme enables systematic examination of interactional structure, including informational contribution, acknowledgement behaviour, clarification moves, and stance marking. The schema, therefore, provides a consistent analytical framework for examining dialogue organisation within the dual-perspective setting.

B. PAIR Dialogue Act Distribution

Figure 4b shows the ten most frequent dialogue acts in PAIR, which together account for the majority of utterances. *STATEMENT* constitutes 51.35% of all acts, followed by *BACKCHANNEL/ACKNOWLEDGE* (9.53%) and *YES ANSWERS* (7.47%). Additional recurring categories include *AGREEMENT/ACCEPT*, *REPEAT-PHRASE*, *HEDGE*, *YES-NO-QUESTION*, and *OPEN-QUESTION*.

The full distribution (Figure 4a) exhibits a long tail of less frequent acts, each contributing fewer than 1% of utterances. These include categories such as *COLLABORATIVE COMPLETION*, *APPRECIATION*, *REJECT*, and *RHETORICAL QUESTION*.

Overall, the distribution reflects a corpus dominated by declarative contributions, supplemented by acknowledgement and inquiry-related acts. This quantitative profile provides the basis for structural analyses discussed in the main text, including examination of participation symmetry and questioning behaviour.

C. PAIR Dialogue Style Analysis

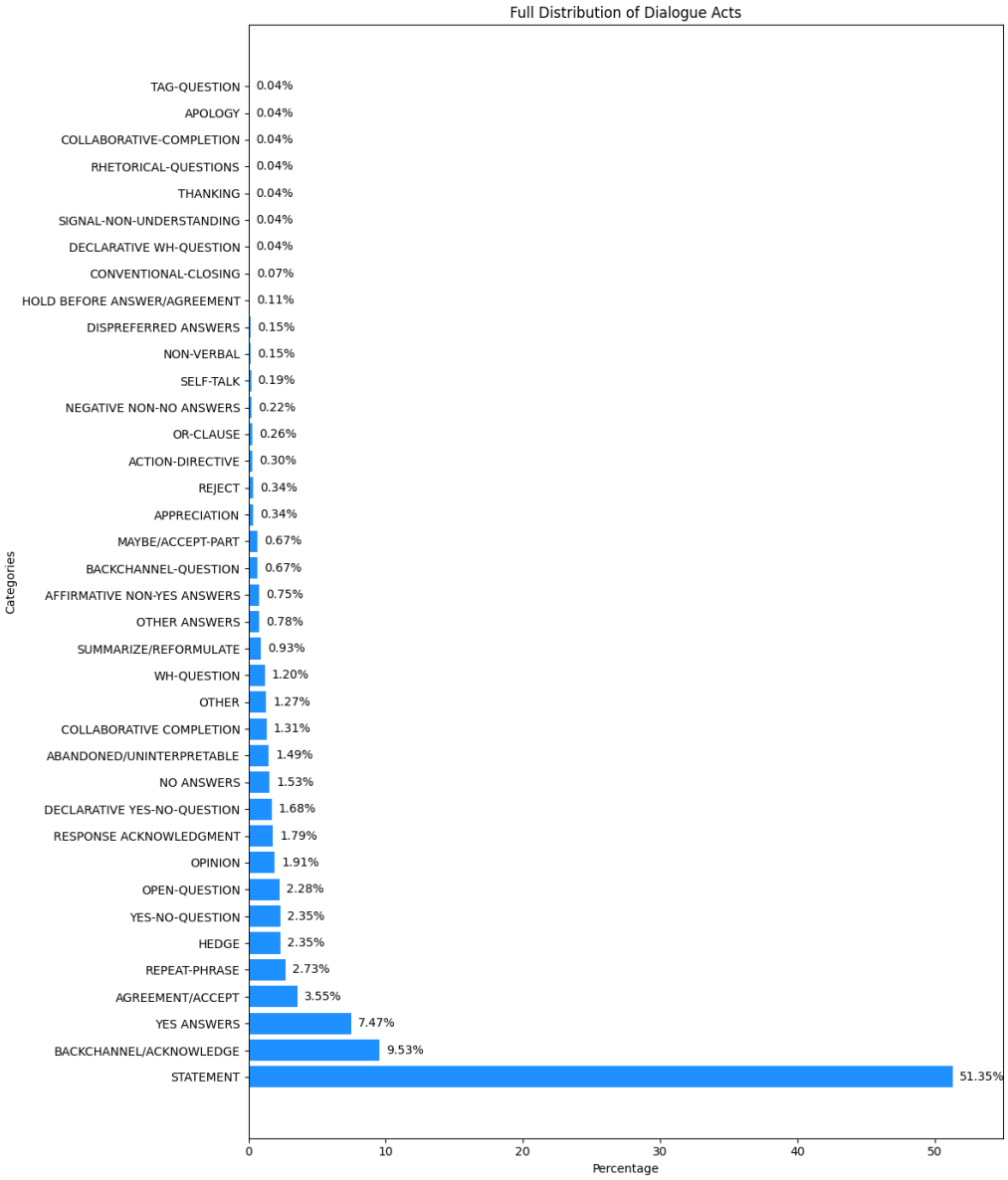
Figure 5 presents the distribution of three aggregated dialogue act categories—*STATEMENT*, *BACKCHANNEL/ACKNOWLEDGE*, and *QUESTIONS*—across all 15 dialogues, organised by scenario and dialogue instance. Each subplot compares the contributions of Speakers 00 and 01.

Across dialogues, *STATEMENT* acts constitute the largest proportion of contributions, with some sessions showing clear asymmetry in statement production between speakers. *BACKCHANNEL/ACKNOWLEDGE* acts occur consistently across sessions, though at lower frequencies. *QUESTION* acts vary more substantially between dialogues, with some sessions exhibiting concentrated questioning behaviour by one speaker.

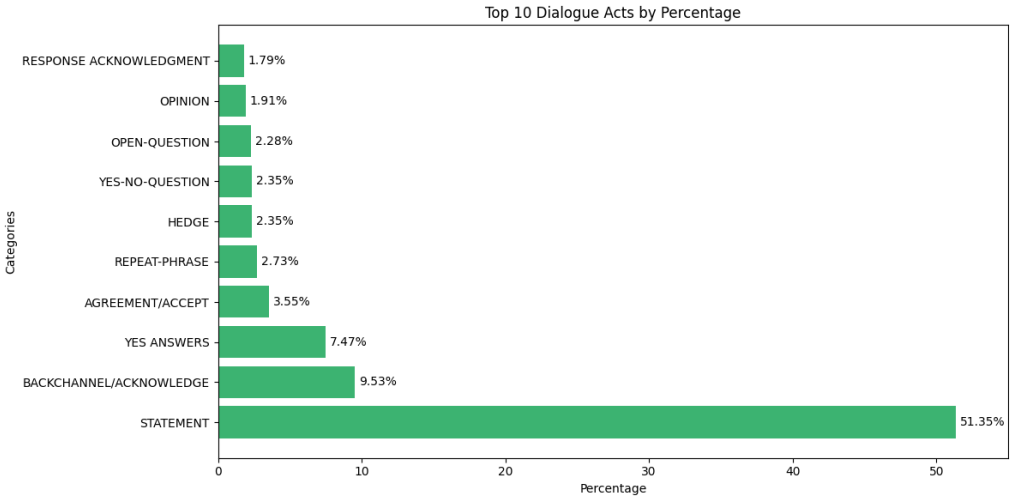
These distributions illustrate variation in speaker contributions and dialogue-act composition across sessions. The figure complements the quantitative metrics presented in the main analysis by providing a dialogue-level view of participation symmetry and act concentration.

Table 3: The 42 dialogue act labels from Stolcke et al. (2000) and examples from the PAIR Dataset. Examples with - - - indicate no presence in the dataset.

Tag	Example
STATEMENT	He washed it.
BACKCHANNEL/ACKNOWLEDGE	Mhmm.
OPINION	I think it didn't look related.
ABANDONED/UNINTERPRETABLE	He looked...
AGREEMENT/ACCEPT	That's right.
APPRECIATION	That's a good way.
YES-NO-QUESTION	Is that the guy in the white T-shirt?
NON-VERBAL	hahaha
YES ANSWERS	Yes.
CONVENTIONAL-CLOSING	I can't think of anything else either.
WH-QUESTION	What else was there?
NO ANSWERS	No.
RESPONSE ACKNOWLEDGMENT	Ah okay.
HEDGE	I don't know, does that make sense?
DECLARATIVE YES-NO-QUESTION	Is there a boiler as well, I think, or a sink?
OTHER	No, he could have bought the T-shirt at a charity shop.
BACKCHANNEL-QUESTION	okay oh you saw his shoes?
QUOTATION	- - -
SUMMARIZE/REFORMULATE	yeah, you mean reuse them.
AFFIRMATIVE NON-YES ANSWERS	He may well have.
ACTION-DIRECTIVE	Let's draw.
COLLABORATIVE COMPLETION	and the boy came to get it. Yeah.
REPEAT-PHRASE	Yeah, messy.
OPEN-QUESTION	Anything else?
RHETORICAL-QUESTIONS	there must be what 30 bikes if not more?
HOLD BEFORE ANSWER/AGREEMENT	Hold on, What was the... What was the take thing?
REJECT	No, it wasn't.
NEGATIVE NON-NO ANSWERS	I don't think so.
SIGNAL-NON-UNDERSTANDING	What?
OTHER ANSWERS	I don't know
CONVENTIONAL-OPENING	- - -
OR-CLAUSE	or was that the T-shirt the guy was wearing?
DISPREFERRED ANSWERS	Something like that.
3RD-PARTY-TALK	- - -
OFFERS, OPTIONS & COMMITS	- - -
SELF-TALK	I'm trying to think.
DOWNPLAYER	- - -
MAYBE/ACCEPT-PART	Could be
TAG-QUESTION	You know, based on how it looked, right?
DECLARATIVE WH-QUESTION	- - -
APOLOGY	- - -
THANKING	Thank you.



(a) Distribution of Dialogue Acts in PAIR Corpus



(b) Top Ten Dialogue Acts by Percentage

Figure 4: Dialogue Acts Statistics

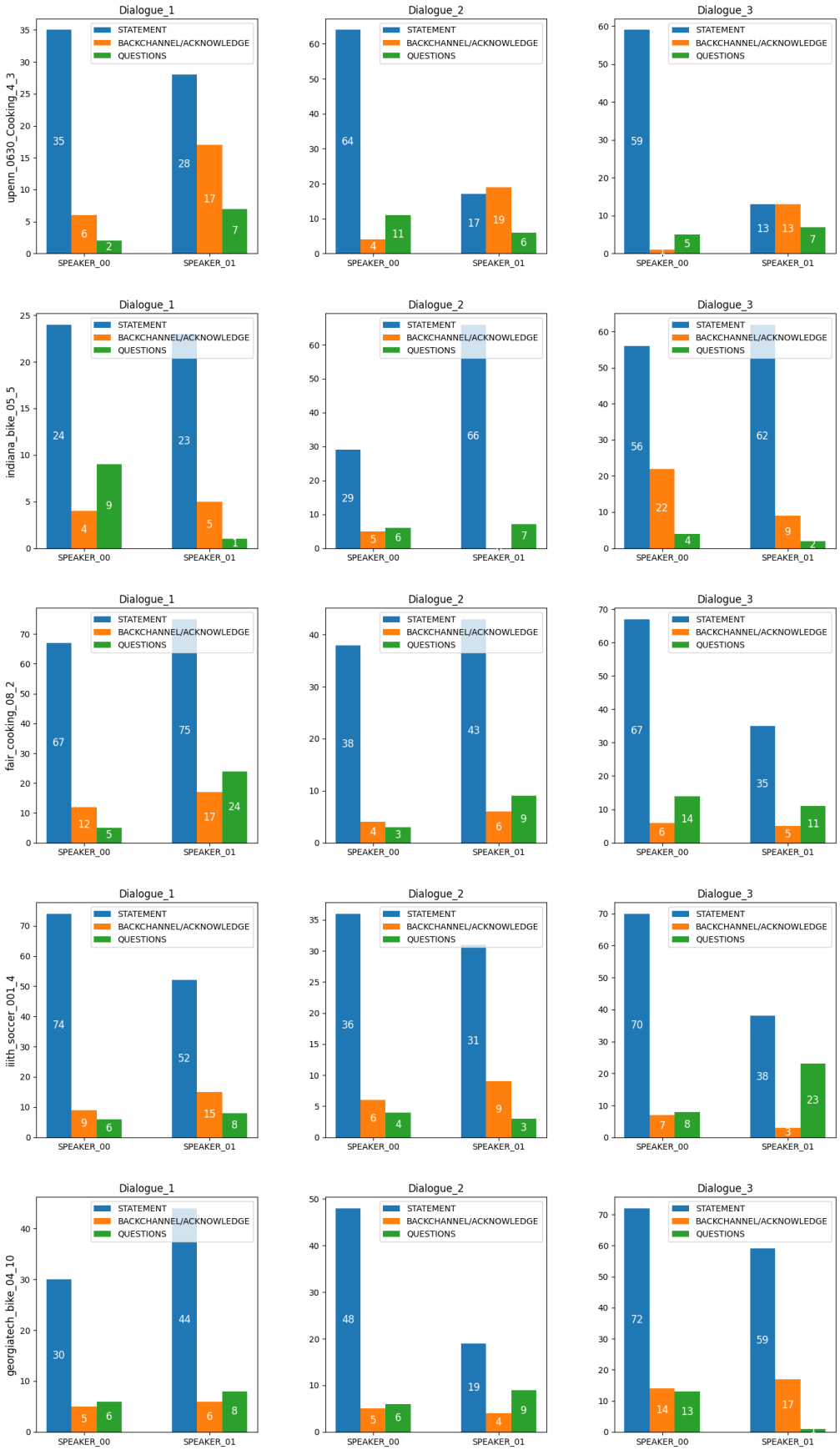


Figure 5: Analysis of Statement, Acknowledge, and Question for dialogue style identification. Where 'Questions' is a grouped DA category consisting of the following DAs: YES-NO-QUESTION, WH-QUESTION, DECLARATIVE YES-NO QUESTION, BACKCHANNEL-QUESTION, OPEN-QUESTION, RHETORICAL-QUESTIONS, OR-CLAUSE, and TAG-QUESTION.