JMDC-MM：「マーダーミステリー」における日本語多人数議論対話コーパス構築の試み

山本ゆのか[1] 西尾琉惺[1] 松垣颯夏[1] 鈴木丈慈[2]
黒古歩希[2] 大橋玲音[2] 坪倉和哉[2] 小林邦和[1,2]
[1]愛知県立大学 情報科学部 [2]愛知県立大学大学院 情報科学研究科
is{241094,241061,241080}@cis.aichi-pu.ac.jp im{251007,251004,251003}@cis.aichi-pu.ac.jp
id231001@cis.aichi-pu.ac.jp kobayashi@ist.aichi-pu.ac.jp

概要

近年，大規模言語モデルの発展に伴い，多人数が関与する複雑な状況下での対話コーパスが求められているが，日本語における多人数対話コーパスの整備状況は十分ではない．そこで本研究では多人数議論ゲームである「マーダーミステリー」に着目し，日本語多人数議論対話コーパス JMDC-MM（Japanese Multiparty Discussion Corpus – Murder Mystery）を作成した．大学生・大学院生 16 名，合計 8 セッションのゲームを実施し,合計 2484 発話からなるデータを収集した．コーパスを分析した結果，フェーズごとの議論の盛り上がり方やプレイヤの役割に応じた目標達成度の違いが確認できた．このコーパスは，多人数対話におけるエージェントの役割理解や協調行動のモデル化に関する研究の進展にも寄与すると考えられる．

1 まえがき

近年，大規模言語モデル（Large Language Models: LLM）の発展に伴い，自然言語処理における対話技術も大きく進展している[1]．従来，LLM による対話は主に 1 対 1 の形式で扱われることが多かったが，近年ではより複雑な対話環境，すなわち多人数が関与する多人数対話においても LLM の適用が進められている[2]．そのため，複数のエージェントが相互に推論を行いながら対話を進めるマルチエージェント対話推論において，性能評価を行うための指標が必要とされている．

このような複雑な対話状況における LLM の推論能力を評価するためには，自然で多様な文脈を含む対話データと，それに基づいた体系的なベンチマークが不可欠である．これに対するアプローチの一つとして，議論型ロールプレイゲームである「マーダーミステリー（Murder Mystery）」を活用した研究が注目されている[3][4][5]．マーダーミステリーは，プレイヤがそれぞれ異なる役割や情報をもとに，他者との議論や情報交換を通じて事件の真相に迫るゲームであり，その性質上，複数人の複雑な議論や推論が求められる．したがって，マーダーミステリーは LLM にとっても自然な発話生成・推論能力を評価する上で非常に有効な題材となる. しかしながら，現在までのマーダーミステリーを対象とした研究の多くは英語話者を対象としており[3][4]，日本語における多人数対話，特にマーダーミステリーのような複雑な議論型ロールプレイに関する公開されたデータセットは存在しない．本研究では，前回著者らが作成したデータセット[6]を拡張し，日本語におけるマーダーミステリー議論データの収集と分析を行った．具体的には，4 人 1 組のプレイヤチームによる 90 分間のマーダーミステリーゲーム 2 作品を，計 4 チーム（合計 16 名）に実施し，ゲーム中の自然な議論を対話として記録した．これにより，日本語における多人数で複雑な議論を扱うための新たな対話データセットを構築するとともに，今後のマルチエージェント対話モデルの研究に有益な基盤を提供することを目指す. 本論文の貢献は, 以下の 2 点である．

- マーダーミステリーにおける日本語多人数議論対話コーパス（JMDC-MM）の公開[i]
- マーダーミステリーの議論における発話量と事後アンケートの分析

2 マーダーミステリーの概要

マーダーミステリーとは，参加者が架空の殺人事件の登場人物となり，相互の対話を通じて事件の真

[i] コーパスは以下の URL にて公開した
https://huggingface.co/datasets/Camellia-Dragons/JMDC-MM

相解明を目指す，推理要素を含むロールプレイングゲームである．各参加者（以下，プレイヤ）は，割り当てられた役割（ロール）に基づき，自身の個別目標の達成と犯人の特定を目的として行動する．

ゲームの進行は，一般的に以下の 3 つのフェーズから構成される．

·**導入フェーズ:** プレイヤに役割が割り当てられ，シナリオの背景や各自の公開情報が共有される．

·**議論フェーズ:** プレイヤは，証拠を提示し，相互に尋問を行うことで，他のプレイヤが持つ情報を収集・分析する．

·**推理・投票フェーズ:** 議論内容に基づき，各プレイヤが犯人と思われる人物に投票を行う．投票後，事件の真相が明かされる．

各プレイヤには，役割ごとに「公開情報」と「非公開情報」が与えられる．公開情報は，役柄のプロフィールなど全プレイヤで共有される基本的な設定である．一方，非公開情報には，個人のアリバイ，潜在的な動機，他者との秘密の関係など，そのプレイヤのみが知る情報が含まれる．役割はシナリオによって多様であり，例えば，犯人役には一貫した虚偽証言による容疑の回避が，無実の役には論理的な潔白の証明がそれぞれ求められる．

本ゲームは，「情報の非対称性」という根源的な特性を持つ．各プレイヤが保持する情報は意図的に偏在化されており，他者からの質問への応答などを通じて断片的に開示される．この構造が，プレイヤ間での戦略的な情報操作や信頼関係の構築といった，複雑な社会的インタラクションを誘発する．

さらに，マーダーミステリーは「一回性」という制約を持つ．ゲーム終了時に全ての真相が解明されるため，同一プレイヤが同じシナリオを再体験することはできない．この性質上，本研究のように人間の自然な反応や意思決定プロセスを分析対象とする場合，参加者による初回プレイの記録を収集することが不可欠となる．

3 対話データの収録

3.1 実験の設計

本研究で用いる対話データを収集するため，同一のマーダーミステリーシナリオを用いた実験を実施した．シナリオには，Group SNE 社の「マーダーミステリーミニシリーズ」から「ケイヴァー洞窟の煌めき」および「カナリアは歌わない」を採用した．実験参加者は大学生・大学院生 16 名であり，4 人 1 組のグループを 4 つ構成した. 実験参加者 16 名のうちプレイ回数が 1 回以下の参加者は 14 名で, プレイ回数が 2 回以上の参加者は 2 名であった．実験は，静謐な環境下でテーブルを囲む形式で行い，各セッションは約 90 分間とした．

3.2 実験手順

実験は，(1)事前説明，(2)事前アンケート，(3)ゲームプレイ，(4)事後アンケートの 4 段階で実施した．まず, 参加者に対してゲームの基本ルールを説明し，本研究の目的と手順を伝えた（約 5 分）．次に，性格特性などを測定する事前アンケートに回答を求めた（約 10 分）．その後，参加者は 90 分間のゲームセッションを行った．セッション終了後，ゲーム体験に関する事後アンケートを実施し（約 5 分），全ての実験手続きを終了した．

3.3 収集データ

本実験では，ゲームプレイ中の音声データと，プレイ前後に実施したアンケートデータを収集した．音声データとして，各セッションにおける全対話を IC レコーダーを用いてステレオで録音した．アンケートは，以下の項目から構成される．

·**事前アンケート**: 参加者の個人的特性を測定するため，BigFive 性格特性[7]，コミュニケーション・スキルを測る ENDCORE（簡易版）[8]，特性シャイネス尺度[9]，社会的スキル[10]の 4 項目を調査した．

·**事後アンケート**: ゲーム体験を多角的に評価するため，楽しさ，難易度，疲労度，目標達成度といった主観的評価に加え，役割演技の質や議論への参加度に関する自己評価，他者評価，および参加者のマーダーミステリー経験の有無を調査した．

3.4 音声データの前処理

収録した音声データに対して，以下の 4 段階の前処理を適用した．

(1) フェーズ抽出: 各セッションの録音データから，分析対象である議論フェーズの音声を抽出した．ここで，「ケイヴァー 洞窟の煌めき」の議論フェーズは全 3 回，「カナリアは歌わない」の議論フェーズは全 5 回である．

(2) 音量正規化: 抽出した音声データの音量を，ラウドネス基準である-23 LUFS に統一した．

(3) 話者分離: pyannote.audio[11][12]を用いて，発話区間と話者を自動的に特定した．

(4) 自動音声認識: OpenAI の Whisper large-v3[13]を用いて，分離された発話区間ごとに文字起こしを行った．

3.5 データセットの構成

上記の手続きを経て構築したデータセットは，以下の 3 種類のデータから構成される．

- **発話テキストデータ**: 8 ゲーム（4 グループ×2 作品）ゲーム分の議論フェーズにおける全発話．話者 ID，発話内容，タイムスタンプを含む．
- **アンケートデータ**: 全 16 名分の事前・事後アンケートの回答結果．
- **メタデータ**: 各参加者に割り当てられた役割，ゲームの結果（犯人投票の結果など）．

4 データの分析

本章では構築したデータセットの初期的な分析として，事後アンケートやグループごとの発話量の分析を行う．まず，フェーズごとの議論の盛り上がりを比較するため，フェーズごとに 1 分あたりの発話量を算出した．ここで，連続した同一発話者ラベルを結合して 1 発話とした．1 ゲーム目の各グループのフェーズごとの発話量の折れ線グラフを図 1 に，2 ゲーム目の各グループのフェーズごとの発話量の折れ線グラフを図 2 に示す．

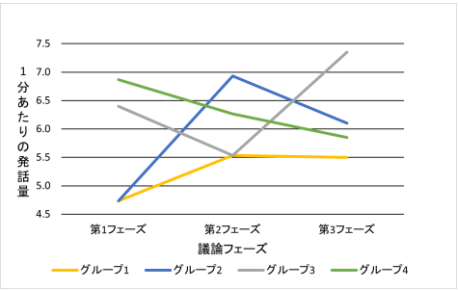

図 1　1 ゲーム目の 1 分あたりの発話量の比較

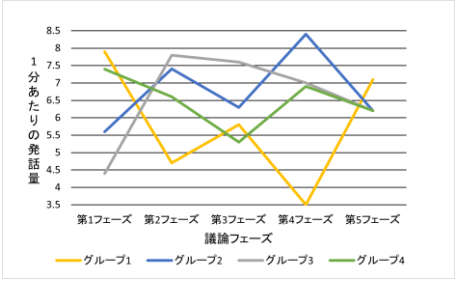

図 2　2 ゲーム目の 1 分あたりの発話量比較

議論フェーズが移行するにつれて発話量が減少するパターンや，最終フェーズに発話量が急激に増加するパターンなど，様々な議論パターンが存在することが分かる．本実験で選定した 2 つのシナリオは，ゲームの開始時点からすべての情報が開示されるのではなく，プレイヤが選択したカードに応じて情報が段階的に開示されていく．そのため，プレイヤが選択したカードに応じて様々な議論展開があり，グループごとに異なる盛り上がり方をしていると考えられる．また，各グループの 1 ゲーム目と 2 ゲーム目の発話量の比較も行った（図 3）．

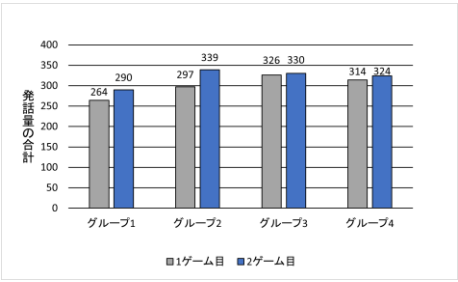

図 3　1，2 ゲーム目の発話量の比較

4 つのグループすべてにおいて，1 ゲーム目より 2 ゲーム目の方が，発話量が増加していることが分かる．これは，プレイヤのプレイ回数が増えたことでプレイヤのゲームに対する理解度が上がり，活発に議論が展開されやすくなった可能性がある．

次に，与えられた役割によって目標達成度に違いが表れるかを分析するため，犯人役と犯人役以外の役の達成度をグラフにまとめた．目標達成度は「達成できた」「だいたい達成できた」「あまり達成できなかった」「達成できなかった」の 4 段階で設定しており，目標達成度に応じて参加者が自己評価を行っている．犯人役の達成度を図 4，犯人役以外の役の達成度を図 5 に示す．

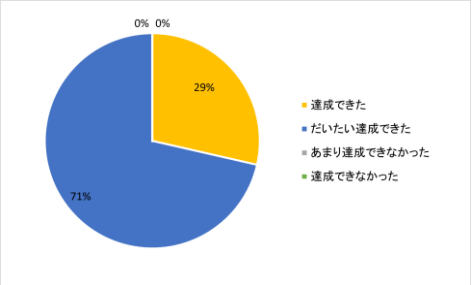

図 4　犯人役に割り当てられた人の目標達成度

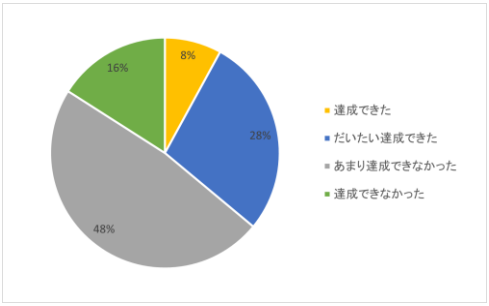

図 5　犯人役以外の役を割り当てられた人の目標達成度

犯人役の 71%が「達成できた」と回答しており，「達成できた」，「だいたい達成できた」と回答した人を合わせると 100%となることが分かる．一方で，犯人役以外の役では 16%が「達成できなかった」と回答しており，「達成できなかった」，「あまり達成できなかった」と回答した人を合わせると 64%となることが分かる．この結果から，本実験で収集したデータでは犯人役の勝率が高いことが分かる．これは，限られた時間では十分に議論が発展せず，犯人が特定されにくくなっているため勝率に偏りが出た可能性がある．本実験はプレイ回数が 1 回以下の参加者が約 87.5%であり，経験の少ないプレイヤが多く，マーダーミステリーの議論に慣れていないことから十分に議論が発展しなかった可能性がある．

さらに，参加者を目標達成度の高い人（「達成できた」または「だいたい達成できた」）と低い人（「あまり達成できなかった」または「達成できなかった」）の 2 群に分割し，事後アンケートの楽しさ，難易度，議論への参加度を比較した．楽しさ，難易度，議論への参加度は 0-100 の範囲で設定しており，参加者が自己評価を行っている．達成度が高い人と低い人を比較した結果を図 6 に示す．

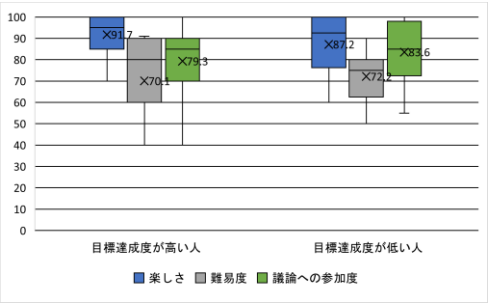

図 6　目標達成度が高い人と低い人の比較

目標達成度が高い人の平均の方が楽しさが約 3.2 ポイント高い結果となった．この結果から，目標達成度が高い方が楽しかったと感じやすいことが分かる．また，目標達成度が低い人の平均の方が難易度が約 2.1 ポイント高い結果となった．この結果から，ゲームの難易度が目標達成度を決めていることが分かる．また，達成度が低い人の方が議論への参加度が約 4.3 ポイント高い結果となった．議論への参加度が高い人は，自分の非公開情報を公開する，または他のプレイヤの非公開情報を引き出そうとする．前者は，自分の非公開情報を多く公開しているため，目標達成度が下がる．後者は，他のプレイヤと対立関係を生む可能性がある．そのため，議論への参加度が高い人は目標達成度が低下すると考えられる．ただし，本実験の参加者は経験の少ないプレイヤが多いため，経験の多いプレイヤにおいてこのような傾向が観察されるとは限らない．

以上の分析結果から，フェーズごとの議論の盛り上がり方やプレイヤの役割に応じた目標達成度の違いが確認された．これらの結果は，マーダーミステリーにおけるプレイヤの発話行動をモデル化するための情報を提供するものであり将来的には本研究で公開するデータセットと合わせて，多人数議論対話における AI エージェントの実現に貢献できるものと思われる．また，プレイヤの熟練度による行動の差異に着目して分析を進めることで，初心者が陥りやすい誤った判断傾向や発話の特性を整理することが可能であると考えられる．そのため，マーダーミステリーの初心者支援システムの設計にも寄与する可能性がある．

5 むすび

本研究では，日本語における多人数対話データの不足という課題に着目し，議論型ロールプレイゲーム「マーダーミステリー」を用いた日本語多人数議論対話コーパス JMDC-MM を構築して公開した．このコーパスは，多人数対話におけるエージェントの役割理解や協調行動のモデル化に関する研究の進展にも寄与すると考えられる．本実験では，基礎的な分析にとどまっており，またデータも少量であるため，データの拡充および議論の質や事前・事後アンケートとの関連の分析を行うことが今後の課題となる．より多様なシナリオや参加者属性を含むデータの拡充を通じて，マルチエージェント対話における言語モデルの能力を多角的に評価・分析する枠組みの構築を目指す．

謝辞

本研究の一部は公益財団法人日東学術振興財団の助成を受けている．

参考文献

[1] Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang and Kam-Fai Wong. A survey of the evolution of language model-based dialogue systems. arXiv preprint arXiv:2311.16789, 2023.

[2] Sagar Sapkota, Mohammad Saqib Hasan, Mubarak Shah and Santu Karmaker. Multi-Party Conversational Agents: A Survey. arXiv preprint arXiv:2505.18845, 2025.

[3] Qinglin Zhu, Runcong Zhao, Bin Liang, Jinhua Du, Lin Gui and Yulan He. Player*: Enhancing llm-based multi-agent communication and interaction in murder mystery games. arXiv preprint arXiv:2404.17662, 2024.

[4] Yin Cai, Zhouhong Gu, Zhaohan Du, Zheyu Ye, Shaosheng Cao, Yiqian Xu, Hongwei Feng and Ping Chen. MIRAGE: Exploring How Large Language Models Perform in Complex Social Interactive Environments. arXiv preprint arXiv:2501.01652, 2025.

[5] Nonomura Ryota and Mori Hiroki. Who speaks next? Multi-party AI discussion leveraging the systematics of turn-taking in Murder Mystery games. Frontiers in Artificial Intelligence, 2025, 8: 1582287.

[6] 西尾 琉惺，山本 ゆのか，松垣 颯夏，鈴木 丈慈，黒古 歩希，大橋 玲音，坪倉 和哉，小林 邦和．LLM のロールプレイの性能向上に向けたマーダーミステリーの会話データセット構築，第 33 回インテリジェント・システム・シンポジウム(FAN2025), No.FAN-33-012, pp.50-55, 2025.

[7] 小塩真司，阿部晋吾, Pino Cutrone．日本語版 Ten Item Personality Inventory（TIPI-J）作成の試み．パーソナリティ研究，Vol.21，No.1, pp.40–52, 2012.

[8] 藤本学，大坊 郁夫．コミュニケーション・スキルに関する諸因子の階層構造への統合の試み．パーソナリティ研究，Vol. 15，No. 3，pp. 347–361, 2007.

[9] 相川充．特性シャイネス尺度の作成および信頼性と妥当性の検討に関する研究．心理学研究，Vol.62, Issue3, pp.149–155, 1991.

[10] 菊池章夫．思いやりを科学する 向社会的行動の心理とスキル．川島書店，1988.

[11] Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. Proc. INTERSPEECH 2023, 2023.

[12] Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. Proc. INTERSPEECH 2023, 2023.

[13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint arXiv:2212.04356, 2022.