# 日本語会話における相互行為プロファイルの再現可能計測

## タイミング・フィラー・応答型と監査可能 LLM

福原 玄

2026 年 2 月 12 日

**概要**

本稿は，日本語会話から抽出される相互行為指標を，**再現可能な計測フレーム**として統合し，(1) 指標群のカバレッジと欠損構造の定量化，(2) 代表指標の分布可視化と要約統計，(3) 統合プロファイルの低次元可視化とクラスタ構造の安定性評価，(4) LLM 解釈の**監査可能性（provenance）**の定量評価，を提示する．500 件の会話 × 話者レコードに対して，PCA により統合プロファイルを可視化し，KMeans（$k = 6$）の seed 反復で ARI 平均 0.541（中央値 0.520，90% 点 0.858）を得た．また LLM 解釈は，used_features 空率 0.010 に対し used_examples 空率 0.412 であり，「参照特徴量の明示」は概ね達成される一方，「根拠例へのリンク」は未完であることが定量的に示された．

# 1 はじめに

会話の「上手い／下手」を主観評価や粗い類型に還元するのではなく，会話の中で生じる相互行為（間・言い淀み・応答の型）を**測定**として定義し，再現可能に計算することは，会話研究・応用（支援，評価，システム設計）において重要である．本研究の狙いは，複数の相互行為指標を単一フレームで統合し，**(i) 観測可能性（カバレッジ／欠損）**と **(ii) 統合プロファイル（低次元表現／クラスタ安定性）**を同時に議論できる形に整える点にある．さらに，LLM を用いた解釈を「正しさ」ではなく**監査可能性**として定量化し，根拠に辿れる設計へ接続する．

# 2 方法

## 2.1 データと解析単位

解析対象は，JSONL 形式で保存された 500 行のレコードである．各レコードは主として会話 × 話者単位の集計値を含む（出力スキーマは機械的に確定）．

## 2.2 指標群（feature families）

本稿で扱う指標群は以下である：

- **PG（Timing）**：総時間，発話時間，発話率など，タイミング情報に基づく指標群.
- **FILL（Disfluency）**：フィラー（例：えっと，えー等）の回数や正規化指標.
- **RESP（Response-typing）**：特定の投げ方条件（例：「ね」「よ」）直後の相槌率や，応答先頭語分布のエントロピー（多様性）.
- **CL（Derived）**：統合プロファイルから導出される PCA 座標等.

本スナップショットでは IX 群（修復・話題連結などに相当する別指標群）は同梱されていないため，統合解析は PG/FILL/RESP（＋派生 CL）に限定する.

## 2.3 カバレッジと欠損の扱い

列の存在／非存在は「カバレッジ」として分離し，存在する列における NA を欠損として扱う．欠損には，(i) 入力情報の欠如（例：タイミング情報なし），(ii) 条件付き指標の分母 0，が含まれる．欠損率は指標群ごとに集計し報告する.

## 2.4 統合プロファイルの構築と標準化

PG/FILL/RESP を連結して統合プロファイル行列を構成し，(1) 非欠損数が十分（例：30 以上），(2) 変動が十分（ユニーク値 3 以上）な列のみを採用した．欠損値は列中央値で補完し，列ごとに z-score 標準化した.

## 2.5 PCA とクラスタ安定性

標準化済み行列に PCA を適用し 2 次元へ射影した．クラスタ構造は KMeans（$k = 6$）で求め，seed を変えた反復により割当の一致度（ARI）を評価した.

## 2.6 LLM 解釈の監査可能性

LLM 出力の監査可能性を，(i) used_features 空率，(ii) used_examples 空率として定量化した．これは「内容の正しさ」ではなく，「根拠に辿れる形で出力されているか」を測る.

表 1: 指標群のカバレッジと欠損率（平均）

| 指標群 | 検出列数 | 欠損率（平均） |
|---|---|---|
| PG | 18 | 0.344 |
| FILL | 13 | 0.000 |
| RESP | 3 | 0.040 |
| CL | 3 | 0.260 |
| IX | 0 | – |

表 2: 代表指標の要約統計

| 指標 | $N$ | mean | p50 | p90 |
|---|---|---|---|---|
| RESP_NE_AIZUCHI_RATE | 480 | 0.576 | 0.583 | 0.808 |
| RESP_NE_ENTROPY | 480 | 3.185 | 3.278 | 4.264 |
| RESP_YO_ENTROPY | 480 | 2.154 | 2.322 | 3.459 |
| FILL_cnt_total | 500 | 59.118 | 42.000 | 130.000 |
| FILL_has_any | 500 | 46.066 | 33.000 | 93.000 |
| PG_speech_ratio | 371 | 0.327 | 0.319 | 0.522 |

# 3 結果

## 3.1 カバレッジと欠損

検出された列数は合計 89 であり, 内訳は PG=18, FILL=13, IX=0, RESP=3, CL=3 であった. 欠損率の平均は指標群ごとに異なり，CL mean=0.260，FILL mean=0.000，PG mean=0.344, RESP mean=0.040 であった．PG および CL の欠損は，タイミング情報が付与されないレコードが一定割合存在することを示唆する.

## 3.2 代表指標の分布と要約統計

代表指標の要約統計を表 2 に示す．RESP では,「ね」直後の相槌率の平均は 0.576（$N = 480$), 中央値 0.583，90% 点 0.808 であった．応答先頭語の多様性（entropy）は,「ね」条件で平均 3.185（中央値 3.278，90% 点 4.264),「よ」条件で平均 2.154（中央値 2.322，90% 点 3.459）であり，条件により分布特性が異なることが示された．FILL は欠損がなく，FILL_cnt_total の平均は 59.118（中央値 42，90% 点 130，$N = 500$）であった．PG_speech_ratio はタイミング情報があるサブセットで $N = 371$，平均 0.327（中央値 0.319，90% 点 0.522）であった.

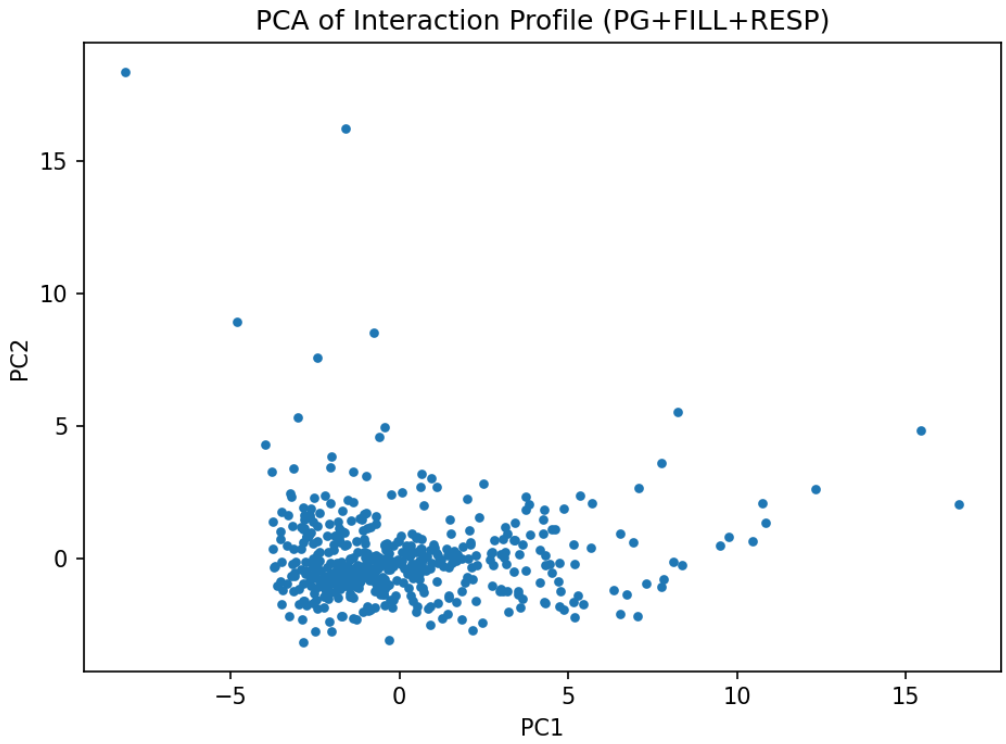



図 1: 統合相互行為プロファイル（PG+FILL+RESP）の PCA 散布図

表 3: クラスタ安定性（ARI）

| $k$ | repeats | mean | p50 | p90 |
|---|---|---|---|---|
| 6 | 20 | 0.541 | 0.520 | 0.858 |

## 3.3 統合プロファイルとクラスタ安定性

PG/FILL/RESP を統合した相互行為プロファイルを標準化し，PCA で 2 次元に射影した（図 1）.

クラスタリングの再現性（seed 違い）を ARI で評価したところ，$k = 6$，repeats=20 で，ARI mean=0.541（p50=0.520，p90=0.858）であった．これは，**一定程度安定なクラスタ構造が存在する一方で**，初期条件により分割が揺らぐ領域もあることを示す.

## 3.4 LLM 解釈の監査可能性

LLM 解釈の監査可能性を provenance（/）の空率で定量化した．全ラベル数は 931 であり，used_features 空率は 0.010 と低かった一方，used_examples 空率は 0.412 と高く，**根拠例へのリンクが未完**であることが示された.

表 4: LLM 解釈の監査可能性（provenance）

| 指標 | 値 |
|---|---|
| ラベル総数（n_labels） | 931 |
| `used_features` 空率 | 0.010 |
| `used_examples` 空率 | 0.412 |

## 4 考察

本稿は，相互行為指標を統合し，低次元空間での分布可視化とクラスタ安定性評価を提示した．ARI の分布（mean 0.541，p90 0.858）は，安定な分割が成立するサブ構造がある一方で，境界が曖昧な領域では初期条件により割当が変動し得ることを示唆する．この性質は，統合プロファイルを「離散的なタイプ」に固定するのではなく，連続空間（PCA 座標）として扱う設計と相性が良い．

また LLM 解釈について，used_features の空率が低いことは，「説明が参照した指標の明示」が一定達成されていることを意味する．一方，used_examples 空率 0.412 は，説明が具体例へ接続されない割合が大きいことを示し，**監査可能性のボトルネックが「例リンク」側にある**ことが定量的に明確になった．今後は，プロンプト制約（引用 ID の必須化）と UI（引用リンクの強制表示）により，解釈を根拠例へ一貫して接続することが重要である．

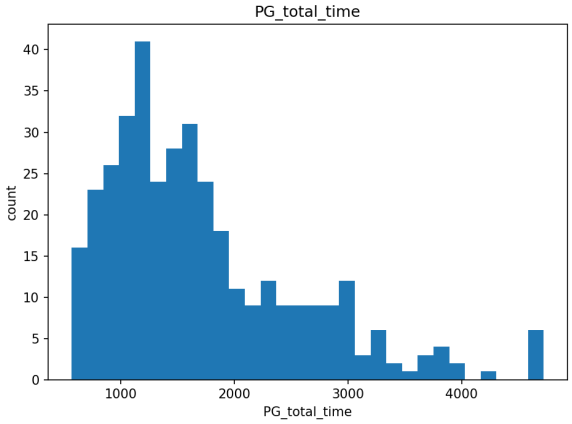
## 5 限界と今後の課題

第一に，本スナップショットでは IX 群（修復・話題連結などに相当する指標群）が同梱されていないため，統合解析は PG/FILL/RESP に限定された．今後は，別テーブルに存在する指標群の統合を行い，統合プロファイルの説明力と安定性の変化を評価する．第二に，欠損補完（中央値）・列選別閾値などの設計は再現性を重視した固定規則であり，最適性は今後の比較検証（代替補完，ロバスト PCA 等）で評価する必要がある．第三に，LLM 解釈は監査可能性を測る段階であり，内容の妥当性評価は別途設計する．

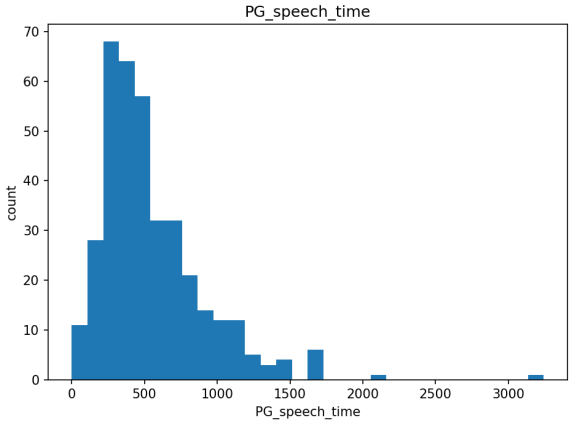
## 6 結論

日本語会話の相互行為指標を再現可能な形で統合し，カバレッジ／欠損構造，代表指標の分布，統合プロファイルの可視化とクラスタ安定性，および LLM 解釈の監査可能性（provenance）を定量的に報告した．特に，used_examples の欠落が監査可能性の主要課題であることを明確化した点は，「解釈を科学っぽく強くする」ための直接的な改善指針となる．

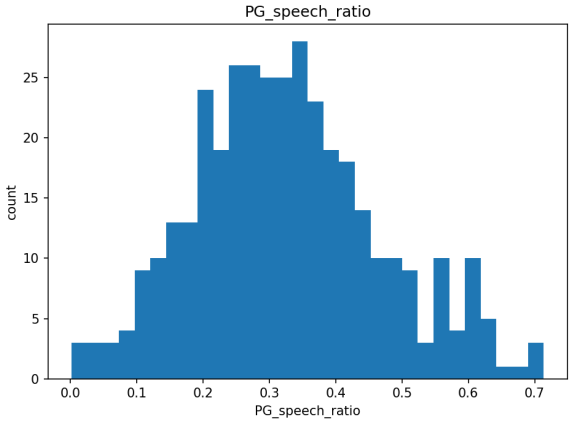
## 付録 A　図：代表ヒストグラム（例）



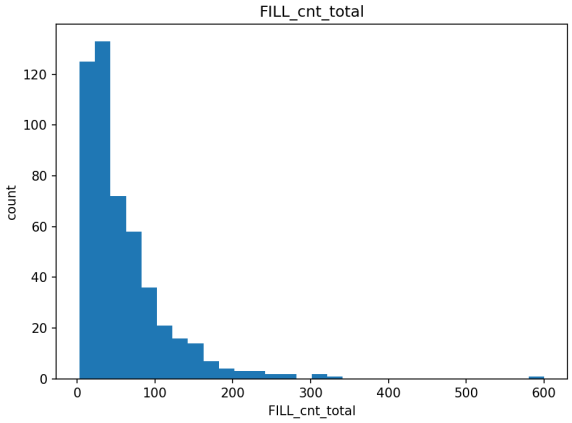

(a) PG_total_time



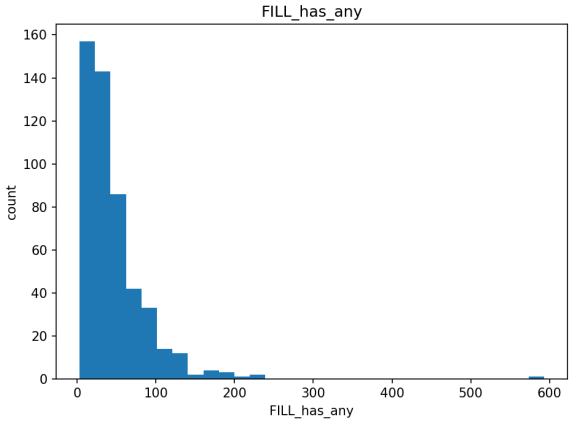

(b) PG_speech_time



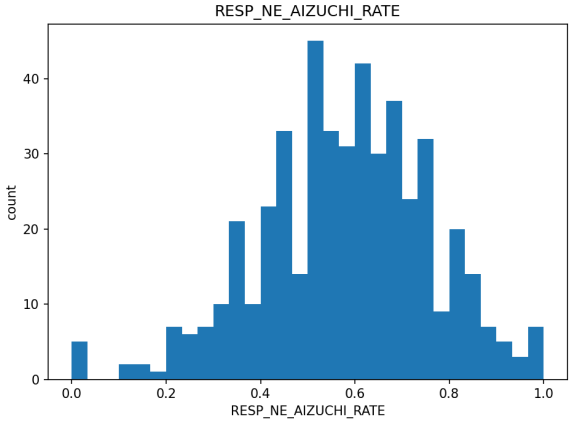

(c) PG_speech_ratio



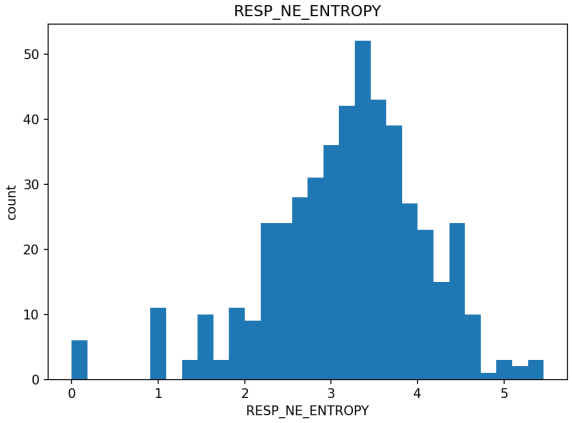

(d) FILL_cnt_total



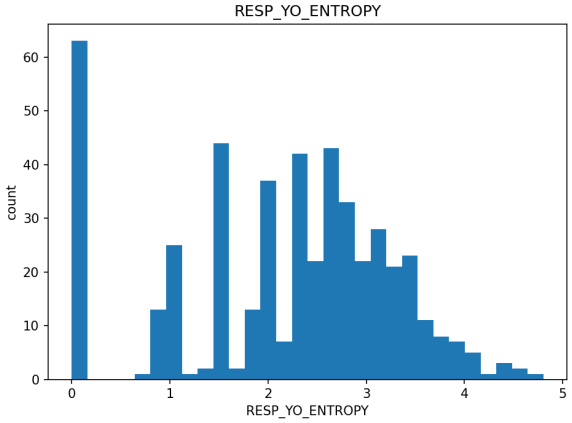

(e) FILL_has_any




(f) RESP_NE_AIZUCHI_RATE




(g) RESP_NE_ENTROPY




(h) RESP_YO_ENTROPY

図 2: 代表指標の経験分布（ヒストグラム）