日本語日常会話における相互行為特徴量の定量化指標の提案
——コーパス基本情報および Big Five 性格特性による妥当性検証——

福原 玄 * 山下 祐一 † 宗田 卓史 ‡

2026 年 4 月 6 日

概要

本研究は，日本語日常会話コーパス（CEJC）から抽出可能な 19 の相互行為特徴量——発話タイミング（PG），フィラー（FILL），相互行為構造（IX），応答型（RESP）——を定量化指標として提案し，その妥当性をコーパス基本情報および Big Five 性格特性との関連から検証する．CEJC home2 サブセットの高品質会話 120 件（会話 × 話者ペア）を対象に，まず 19 特徴量の分布特性とカテゴリ内・カテゴリ間の相関構造を記述した．次に，コーパス基本情報（性別・年齢）との関連分析により，発話率やフィラー使用率が性別・年齢と有意に関連することを確認し，特徴量の表面的妥当性を示した．さらに，4 つの大規模言語モデル（LLM）が推定した Big Five スコアを外部基準として Ridge 回帰（5-fold CV, subject-wise split）と置換検定（5000 回）を適用した結果，Conscientiousness（C: 誠実性）は 4 教師中 3 教師で有意な予測精度を示し（$r = 0.390$–0.447, $p < 0.002$），教師間一致度も $\bar{r} = 0.699$ と最も高く，教師横断で最も頑健な予測対象であった．アンサンブル Big5 では Agreeableness（A）が最高の $r = 0.465$ を示したが，A の教師間一致度は $\bar{r} = 0.435$ と低く，教師依存性が大きい．Permutation 回帰係数検定および Bootstrap 分散ベース安定性分析により，発話率（PG_speech_ratio），沈黙系 3 指標（PG_pause_mean / p50 / p90），YES/NO 応答率（IX_yesno_rate / IX_yesno_after_question_rate），「ね」直後応答多様性（RESP_NE_ENTROPY）が C の主要な関連因子として同定された．本研究は，「性格推定モデルの構築」ではなく「日本語会話における相互行為特徴量の定量化指標の提案」として位置づけられ，日本語会話コーパスにおける初の体系的な特徴量提案・検証研究である．

* 筆頭著者．合同会社 Lead lea（リーレア）．連絡先: genfukuhara@leadlea.com
† 国立精神・神経医療研究センター（NCNP）神経研究所 疾病研究第七部
‡ 国立精神・神経医療研究センター（NCNP）

1　はじめに

会話における相互行為——発話のタイミング，言い淀み，応答の型，修復——は，話者の社会的・認知的特性を反映する重要な手がかりである．会話分析（Conversation Analysis）の伝統では，ターンテイキング，修復連鎖，隣接ペアといった現象が質的に精緻に記述されてきた．しかし，これらの知見を大規模コーパスに対して定量的・再現可能に計測する手法は十分に確立されていない．談話分析や NLP の分野では，発話タイミングやフィラー使用率といった音響・語彙レベルの指標が個別に用いられてきたものの，会話の相互行為構造を体系的に捉える指標セットの提案は限られている．

近年，大規模言語モデル（LLM）の急速な発展に伴い，会話テキストから臨床指標や心理特性を直接推定する試みが増加している．Hu ら [1] は，ADOS-2 の対話転記からゼロショット LLM を用いて臨床項目の有無を推定し，Altozano ら [2] は保護者の自由回答から LLM による ASD/TD 分類を試みた．Mun ら [3] は韓国語の子ども発話から PLM を用いた重症度回帰を行い，国内では Nakamura ら [4] が ADOS-2 会話テキストから BERT 由来の特徴量と LightGBM による分類を報告している．これらの研究は LLM/PLM の高い推定能力を示す一方で，共通する課題として**説明可能性（explainability）**の欠如が挙げられる．すなわち，LLM がどのような会話行動に基づいて推定を行っているかが不透明であり，臨床応用や研究の再現性の観点から，推定結果を人間が解釈可能な特徴量で裏付ける必要性が高まっている．

この課題に対して，本研究では会話の相互行為を定量化する特徴量を 2 つの群に整理して提案する．第一群は**既存研究ベースの特徴量（Classical Features, 10 個）**であり，発話タイミング系 8 変数（PG: 発話率，沈黙の長さ，応答遅れ，重なり率）とフィラー系 2 変数（FILL: フィラー出現率，100 文字あたりフィラー率）から構成される．これらは音声研究・談話分析・NLP において広く用いられてきた指標であり，先行研究との比較可能性を担保する．第二群は**新規提案の特徴量（Novel Features, 9 個）**であり，相互行為構造系 5 変数（IX: 修復開始率，YES/NO 応答率，語彙重なり）と応答型系 3 変数（RESP: 終助詞直後の相槌率，応答エントロピー），および沈黙変動性 1 変数（PG_pause_variability）から構成される．これらは会話分析・相互行為論の知見——修復連鎖の組織化，隣接ペアの選好構造，終助詞の語用論的機能——を踏まえて新たに定義した指標であり，本研究の新規性の核をなす．

本研究が採用するアプローチは，LLM と古典的特徴量の**相互補完**にある．具体的には，(1) LLM を「仮想教師」として会話テキストから Big Five 性格スコアを推定し，(2) その推定値を外部基準として，古典的特徴量（Classical）および新規特徴量（Novel）による回帰分析で解釈可能な予測因子を同定する．さらに，(3) 回帰分析で同定された寄与特徴量が会話分析の知見と整合するかを検討し，(4) 新規特徴量の追加が予測精度を向上させるかをベースライン比較で検証する．この「LLM で推定→古典特徴量で解釈→新たな特徴量を定義→ LLM で検証」という循環的な枠組みは，LLM の推定能力と古典的指標の説明可能性を相互に活かす研究設計として位置づけられる．

以上を踏まえ，本研究は日本語日常会話コーパス（CEJC）を対象に，Classical 10 個 + Novel

9 個 = 計 19 の相互行為特徴量を定量化指標として提案し，その妥当性を複数の外部基準から検証する．**重要な点として，本研究の主たる貢献は特徴量の提案そのものにあり，**Big Five 性格特性との関連分析は，提案指標の構成概念妥当性（construct validity）を検証するための手段として位置づけられる．

提案特徴量の妥当性は，以下の 2 段階で検証する．第一段階（1°：提案特徴量の抽出）として，19 特徴量の分布特性を記述し（3.1 節），カテゴリ内・カテゴリ間の相関構造を分析する（3.2 節）．第二段階（2°：外部指標を用いた妥当性検証）として，コーパス基本情報（性別・年齢）との関連から表面的妥当性を確認し（3.3 節），さらに 4 つの LLM 教師のアンサンブル Big Five スコアを主たる外部基準として構成概念妥当性を検証する（3.4 節）．Big5 との関連分析では，subject-wise split によるリーク防止，置換検定（5000 回）による統計的有意性の検証，Bootstrap（500 回）による係数安定性の評価，3 段階 Ridge 回帰比較（人口統計のみ→ Classical 特徴量追加→ Novel 特徴量追加）による段階的な追加効果の検証，および複数 LLM 教師間の一致度による頑健性検証を組み合わせた包括的な評価設計を採用する．

2 方法

2.1 データ

2.1.1 コーパスと前処理パイプライン

本研究では，日本語日常会話コーパス（Corpus of Everyday Japanese Conversation; CEJC）[*1] の home2 サブセット（自宅・少人数会話）を使用した．home2 サブセットを選択した理由は，(1) 自宅という比較的リラックスした環境で収録されており話者の自然な会話行動が反映されやすいこと，(2) 2 者間会話が中心であり相互行為特徴量の計算が明確に定義できること，による．

CEJC home2 サブセットから分析対象データを構築するパイプラインは以下の通りである．まず，home2 に含まれる全会話ファイルから，各会話の各話者について発話タイミング情報（発話開始・終了時刻）と転記テキストを抽出した．次に，本研究独自の品質フィルタ（HQ1）を適用し，分析に十分な品質を持つ会話 × 話者ペアを選定した．最終的に $N = 120$ の会話 × 話者ペア（以下「レコード」と呼ぶ）を分析対象とした．

2.1.2 品質フィルタ（HQ1）

HQ1 フィルタは本研究で独自に定義した品質基準であり，以下の条件を満たすレコードのみを採用する: (1) 発話タイミング情報（発話開始・終了時刻）が完全に付与されていること（タイミング情報が欠損している会話は，沈黙・応答遅れ等の PG 系特徴量が計算できないため除外），(2) 当該話者の発話ペア数（前発話と応答のペア）が十分であること（発話ペア数が極端に少ない場合，IX 系・RESP 系特徴量の推定が不安定になるため除外）．これらの基準は，先行研究で一般的に用

[*1] CEJC は国立国語研究所が構築・公開する日本語日常会話の大規模コーパスであり，自宅・職場・公共空間など多様な場面の自然会話を収録している．

いられるフィルタではなく，本研究の特徴量計算パイプラインに固有の要件に基づいて設計したものである．

2.1.3 CEJC メタ情報の紐付け

CEJC では，各会話に一意の会話 ID（conversation_id，例: K001_011）が，各話者に一意の話者 ID（cejc_person_id，例: K001）が付与されている．1 つの会話には複数の話者が参加するため，分析の基本単位は conversation_id × cejc_person_id のペアとなる．

話者の属性情報（性別，年齢帯，出身地，居住地，職業等）は CEJC が提供するメタデータファイル（話者.csv）に格納されており，cejc_person_id をキーとして紐付けられる．また，各話者がどの会話に参加しているかの対応関係は話者会話対応表（話者会話対応表.csv）に記録されており，conversation_id × cejc_person_id の組み合わせで参照できる．本研究では，これらのメタデータを統合し，$N = 120$ レコード全件について話者属性の紐付けに成功した．

2.1.4 話者の重複と性別内訳

$N = 120$ レコードは 74 名のユニーク話者から構成される．CEJC では同一話者が複数の会話に参加している場合があり，本データセットでも 25 名の話者が 2 件以上の会話に参加しており，これらの重複話者に属するレコードは 71 件（全体の 59.2%）に達する．残りの 49 名は各 1 件の会話にのみ参加している．ユニーク話者 74 名の性別内訳は，女性 38 名・男性 36 名であった[*2]．

この話者重複は，回帰分析における交差検証の設計に影響を与える．同一話者が訓練データとテストデータの両方に含まれると，話者固有の特徴量パターンを通じた情報リークが生じうる．本研究では，この問題に対処するため，交差検証において同一 cejc_person_id のレコードが訓練セットとテストセットに分割されないよう制御する subject-wise split を採用した（2.4 節参照）．これにより，重複話者を含むデータセットにおいても，予測精度の評価が話者間の汎化性能を適切に反映するよう設計されている．

各レコードは，1 つの会話における 1 人の話者の相互行為特徴量（2.2 節）と，LLM 教師が推定した Big Five スコア（2.3 節）の組で構成される．

2.2 特徴量

会話の相互行為を定量化するため，19 の特徴量を抽出した．これらの特徴量は，先行研究での使用実績に基づく**既存研究ベースの特徴量（Classical Features, 10 個）**と，会話分析・相互行為論の知見を踏まえて本研究で新たに定義した**新規提案の特徴量（Novel Features, 9 個）**の 2 群に大別される．なお，IX_topic_drift_mean は IX_lex_overlap_mean との完全共線性（$r = -1.00$，定義上 topic_drift $= 1 -$ lex_overlap）のため説明変数から除外し，統制変数として扱った．特徴量数の水増しを避け，査読者からの指摘を未然に防ぐためである．表 1 に各特徴量の名称，カテゴ

[*2] レコード単位では女性 66 件・男性 54 件となる（3.3 節参照）．重複話者に女性が多いため，レコード単位の性別比はユニーク話者の比率と異なる．

リ，分類（Classical/Novel），概要，計算アルゴリズムを示す．

2.2.1 既存研究ベースの特徴量（Classical Features, 10 個）

Classical Features は，音声研究・談話分析・NLP において広く用いられてきた指標であり，発話タイミング系 8 変数（PG）とフィラー系 2 変数（FILL）から構成される．先行研究との比較可能性を担保する役割を持つ．

■PG（Prosodic/Gap: タイミング系，8 変数） 発話タイミングに関する指標は，会話分析におけるターンテイキング研究 [5] の知見を定量化したものであり，音声対話システムや臨床的な会話評価において広く用いられている [7, 8]．発話率（PG_speech_ratio: 発話率）は話者の総発話時間を会話全体時間で除して算出する．沈黙系指標（PG_pause_mean: 平均沈黙長 / PG_pause_p50: 沈黙長中央値 / PG_pause_p90: 沈黙長 90 パーセンタイル）は，同一話者の連続発話間に生じるギャップのうち，閾値（gap_tol = 0.05 秒）以上のものを対象に平均・中央値・90 パーセンタイルを計算する．沈黙の長さは性格特性や認知負荷の指標として先行研究で繰り返し報告されている [9]．応答遅れ系指標（PG_resp_gap_mean: 平均応答遅れ / PG_resp_gap_p50: 応答遅れ中央値 / PG_resp_gap_p90: 応答遅れ 90 パーセンタイル）は，話者交替時の前発話終了から応答開始までのギャップについて同様の統計量を算出する．重なり率（PG_overlap_rate）は，話者交替時のギャップが −gap_tol 秒未満（すなわち重なり）である割合であり，会話における同時発話の頻度を定量化する．Levitan ら [7] は英語対話における重なり率と性格特性の関連を報告しており，本研究ではこれを日本語会話に適用した．

■FILL（Filler: フィラー系，2 変数） フィラー（filled pause）の使用は，談話分析・心理言語学において発話計画過程の指標として広く研究されてきた [10, 11]．本研究では，計画フィラー（planning filler）として「えっと」「えー」「あの」の 3 種のみを採用した．「まあ」「なんか」「こう」「ほら」等は談話標識（discourse marker）としての機能が強く，発話計画過程の指標としてのフィラーとは区別されるため除外した [11]．FILL_has_any（フィラー出現発話数）は，上記 3 種のフィラーを 1 つ以上含む発話の割合である．FILL_rate_per_100chars（100 文字あたりフィラー率）は，フィラー総数をテキスト文字数の 100 分の 1 で除した正規化指標である．

2.2.2 新規提案の特徴量（Novel Features, 9 個）

Novel Features は，会話分析（Conversation Analysis）および相互行為論の知見を踏まえて本研究で新たに定義した指標であり，相互行為構造系 5 変数（IX）と応答型系 3 変数（RESP），および沈黙変動性 1 変数（PG_pause_variability）から構成される．これらは，修復連鎖の組織化 [6]，隣接ペアの選好構造 [5]，終助詞の語用論的機能といった会話分析の理論的枠組みを定量化可能な指標に変換したものであり，本研究の新規性の核をなす．

■IX（Interaction: 相互行為構造系，5 変数） 修復開始（OIR: Other-Initiated Repair）率（IX_oirmarker_rate: 修復開始率）は，OIR マーカーで始まる応答の割合である．本研究で使

表 1　19 特徴量の定義．Class. 列は Classical（既存研究ベース）または Novel（新規提案）を示す．Summary 列に日本語名を併記した．

Name	Cat.	Class.	Summary	Algorithm
PG_speech_ratio	PG	Classical	Speech ratio	Speaker's total speech time / total conversation time. NaN if total_time is 0 or missing.
PG_pause_mean	PG	Classical	Mean pause duration	Mean of intra-speaker consecutive utterance gaps (≥gap_tol sec). NaN if no qualifying gaps.
PG_pause_p50	PG	Classical	Median pause	50th percentile of intra-speaker gaps. NaN if no qualifying gaps.
PG_pause_p90	PG	Classical	90th percentile pause	90th percentile of intra-speaker gaps. NaN if no qualifying gaps.
PG_resp_gap_mean	PG	Classical	Mean response gap	Mean of turn-taking gaps (prev_end → resp_start, ≥gap_tol sec). NaN if no qualifying gaps.
PG_resp_gap_p50	PG	Classical	Median response gap	50th percentile of turn-taking gaps. NaN if no qualifying gaps.
PG_resp_gap_p90	PG	Classical	90th percentile response gap	90th percentile of turn-taking gaps. NaN if no qualifying gaps.
FILL_has_any	FILL	Classical	Filler utterance rate	Proportion of speaker's utterances containing ≥1 filler (etto/ee/ano). NaN if speaker has no utterances.
FILL_rate_per_100chars	FILL	Classical	Filler rate per 100 chars	Total filler count / (text character count / 100). NaN if text_len is 0.
PG_overlap_rate	PG	Classical	Overlap rate	Proportion of turn-taking gaps ¡ -gap_tol (overlaps).
IX_oirmarker_rate	IX	Novel	OIR marker rate	Proportion of responses starting with OIR markers (e?/eQ/nani? etc). Computed over all adjacent pairs where speaker is responder.
IX_oirmarker_after_question_rate	IX	Novel	Post-question OIR rate	OIR marker rate when previous utterance is a question. NaN if no question-preceded pairs.
IX_yesno_rate	IX	Novel	Yes/No response rate	Proportion of responses starting with yes/no prefixes (hai/un/ee/iie etc).
IX_yesno_after_question_rate	IX	Novel	Post-question Yes/No rate	Yes/No rate when previous utterance is a question. NaN if no question-preceded pairs.
IX_lex_overlap_mean	IX	Novel	Lexical overlap	Mean character-bigram Jaccard coefficient between previous utterance and response.
RESP_NE_AIZUCHI_RATE	RESP	Novel	Post-NE aizuchi rate	Proportion of responses that start with aizuchi prefixes when previous utterance ends with NE particle. NaN if n_pairs_after_NE is 0.
RESP_NE_ENTROPY	RESP	Novel	Post-NE response entropy	Shannon entropy of response-initial tokens after NE sentence-final particle. NaN if n_pairs_after_NE is 0.
RESP_YO_ENTROPY	RESP	Novel	Post-YO response entropy	Shannon entropy of response-initial tokens after YO sentence-final particle. NaN if n_pairs_after_YO is 0.
PG_pause_variability	PG	Novel	Pause duration CV	Coefficient of variation (std / mean) of intra-speaker pause durations. NaN if fewer than 2 pauses or mean is 0.

用する OIR マーカーの完全なマッチング文字列リストは以下の通りである: 「え？」「えっ」「えっ？」「なに？」「なに」「は？」「はっ？」「ん？」「んっ？」「へ？」「へっ？」. 会話分析において修復連鎖は相互理解の達成過程を反映する重要な現象であり [6], その定量化は Kendrick[12] や Albert & De Ruiter[13] によって試みられてきた. 本研究では, OIR 率を個人差変数として外部基準（Big Five 性格特性）と関連づける応用が新規であり, また日本語コーパスへの適用も初めてである. IX_oirmarker_after_question_rate（質問直後修復開始率）は, 前発話が質問である場合に限定した OIR 率であり, 隣接ペアの第一部分（質問）に対する修復開始の頻度を捉える. YES/NO 応答率（IX_yesno_rate: YES/NO 応答率, IX_yesno_after_question_rate: 質問直後 YES/NO 率）は, YES/NO プレフィックス（「はい」「うん」「いいえ」等）で始まる応答の割合を, 全応答および質問直後応答について算出する. これらは隣接ペアにおける選好的応答の出現率を定量化した指標である. 語彙重なり（IX_lex_overlap_mean: 語彙重複度）は, 前発話と応答の文字バイグラム Jaccard 係数の平均であり, 話題の連続性を定量的に捉える指標として新たに定義した. Meylan & Gahl[14] は会話レベルでの Jaccard 類似度を報告しているが, 本研究では隣接ペアレベルの文字バイグラム Jaccard を用いる点で測定手法が異なる.

■PG_pause_variability（沈黙長の変動係数，1 変数） PG_pause_variability は, 同一話者の連続発話間に生じる沈黙長の変動係数（CV: 標準偏差/平均）であり, 沈黙パターンの一貫性と変動性を定量化する新規指標である. CV が大きい話者は沈黙長のばらつきが大きく, CV が小さい話者は一定のリズムで発話する傾向を示す. 有効な沈黙が 2 件未満または平均が 0 の場合は NaN とした.

■RESP（Response-typing: 応答型系，3 変数） 応答型系特徴量は, 日本語会話分析の知見——特に終助詞「ね」「よ」の語用論的機能とそれに対する応答パターンの組織化——に基づいて新たに定義した指標である. 終助詞「ね」のマッチングには正規表現 **（よね|だよね|ですよね|だよな|ね）$** を, 終助詞「よ」のマッチングには正規表現 **（だよ|ですよ|よ）$** を使用した. RESP_NE_AIZUCHI_RATE（「ね」直後相槌率）は, 終助詞「ね」で終わる発話の直後に相槌プレフィックスで始まる応答が出現する割合である.「ね」は共感・確認を求める終助詞であり, その直後の相槌率は話者間の共感的応答パターンを反映する. Kita & Ide[15] は「ね」と相槌の関係を質的に記述しているが, 本研究ではこの質的知見を話者レベルの定量指標に変換した点が新規である. RESP_NE_ENTROPY（「ね」直後応答多様性）および RESP_YO_ENTROPY（「よ」直後応答多様性）は, それぞれ「ね」「よ」で終わる発話の直後の応答先頭トークンの Shannon entropy であり, 応答の多様性を反映する. ASD 研究における相槌多様性（backchannel diversity）の Shannon entropy 測定は先行研究で報告されているが [17], 特定の終助詞（ね/よ）を条件とするエントロピー計算は本研究が初めて提案するものである. これらのエントロピー指標は, 特定の終助詞に対する応答の定型性と多様性を定量化する新規の指標である.

■シャノンエントロピーの計算詳細 RESP_NE_ENTROPY および RESP_YO_ENTROPY の計算方法を以下に詳述する. 対象トークンは, 終助詞「ね」（または「よ」）で終わる発話の直後

に出現する応答の先頭 1 文字（ユニグラム）である．各応答先頭文字の出現確率 $p(x)$ を算出し，Shannon entropy H を自然対数（底 e）を用いて以下の式で計算する:

$$H = -\sum_{x} p(x) \ln p(x) \tag{1}$$

ここで，x は応答先頭文字の各タイプ，$p(x)$ はその出現確率（出現回数 / 総応答数）である．H が大きいほど応答の多様性が高く，H が小さいほど応答が定型的であることを示す．なお，対象となる終助詞ペアが存在しない場合（$n\_\text{pairs} = 0$）は NaN とした．

2.2.3 統制変数の除外と欠損値処理

■統制変数の除外 回帰分析においては，会話の長さや発話ペア数に依存する統制変数（EXCL3: n_pairs_total, n_pairs_after_NE, n_pairs_after_YO, IX_n_pairs, IX_n_pairs_after_question, PG_total_time, PG_resp_overlap_rate, FILL_text_len, FILL_cnt_total, FILL_cnt_eto, FILL_cnt_e, FILL_cnt_ano）および IX_topic_drift_mean（IX_lex_overlap_mean との完全共線性 $r = -1.00$ のため）を説明変数から除外し，上記 19 変数のみを用いた．

■欠損値の処理 特徴量の計算において，分母がゼロとなる場合（例: 質問直後の発話ペアが 0 件の場合の IX_oirmarker_after_question_rate）は NaN とした．回帰パイプラインでは，SimpleImputer（strategy=“median”）により中央値で補完した．

19 特徴量の記述統計は 3.1 節（表 3）に示す．

2.3 仮想教師プロトコル

Big Five 性格スコアの目的変数を得るため，4 つの LLM を「仮想教師」として使用した．具体的には，IPIP-NEO-120（International Personality Item Pool, 日本語版 120 項目）の各項目について，会話テキストを参照しながら LLM に 5 件法で回答させ，5 つの性格次元（Openness: O, Conscientiousness: C, Extraversion: E, Agreeableness: A, Neuroticism: N）のスコアを算出した．

使用した 4 モデルは以下の通りである:

- Sonnet4（Anthropic Claude Sonnet 4）
- Qwen3-235B（Alibaba Qwen3, 235B パラメータ）
- DeepSeek-V3（DeepSeek V3）
- GPT-OSS-120B（OpenAI 系オープンソース, 120B パラメータ）

各モデルは AWS Bedrock 経由で推論を実行し，同一の会話 × 話者ペアに対して独立にスコアを付与した．これにより，特定の LLM に依存しない頑健性の検証が可能となる．

使用したプロンプトテンプレートおよび IPIP-NEO-120 の項目例は付録 D に，各 LLM 教師が推定した Big Five スコアの基本統計量は付録 E に，教師間一致度の詳細（5 次元 ×4 教師 ×4 教

師の相関行列）は付録 G にそれぞれ示す．教師間一致度の結果セクションにおける報告は 3.4.6 節を参照されたい．

2.4 回帰モデル

■Ridge 回帰の選択理由 19 の相互行為特徴量を説明変数，LLM 教師が推定した Big Five スコアを目的変数として，Ridge 回帰モデルを構築した．Ridge 回帰（L_2 正則化）を採用した理由は，提案特徴量間に高い相関（多重共線性）が存在するためである．3.2 節で報告するように，同一カテゴリ内の特徴量間には $r = 0.78$–0.97（PG 系沈黙指標間）や $r = -1.00$（IX_lex_overlap_mean と IX_topic_drift_mean の完全共線性）といった高い相関が認められる．通常の最小二乗回帰ではこれらの多重共線性により係数推定が不安定になるが，Ridge 回帰の L_2 正則化により推定の安定性が確保される．ただし，個々の特徴量の係数解釈には，これらの相関構造を考慮する必要がある．

■正則化パラメータの選択 正則化パラメータ α は事前に固定し（$\alpha = 100$），$\alpha \in \{10, 50, 100, 200, 500\}$ の範囲で感度分析を実施した結果，主要な結果（特に Conscientiousness の予測精度および有意性）は α の選択に対して安定していることを確認した（付録 A 参照）．この安定性に基づき，$\alpha = 100$ を採用した．

■交差検証と予測精度指標 5-fold 交差検証（subject-wise split: 同一 cejc_person_id のレコードが訓練セットとテストセットに分割されないよう制御）により予測精度を評価した．予測精度の指標として，各 fold で算出した Pearson 相関係数 r の 5-fold 平均 $\bar{r}$ を用いた．

■アンサンブル Big5 スコア Big5 の目的変数として，4 つの LLM 教師（Sonnet4, Qwen3-235B, DeepSeek-V3, GPT-OSS-120B）の item-level 回答（IPIP-NEO-120 の各 120 項目に対する 5 件法回答）の算術平均から算出したアンサンブル Big5 スコアを主たる外部基準として用いた．アンサンブルを用いる理由は，(1) 特定の LLM 教師に依存しない安定した推定値が得られること，(2) 個別教師ごとの結果一覧による混乱を回避し，単一の推定値に基づく明快な報告が可能となること，(3) 教師間のばらつきが平均化により安定化されることによる．個別教師ごとの結果は付録 F で報告する．

■置換検定（Permutation test） 回帰モデルの統計的有意性を評価するため，置換検定を実施した．目的変数を 5000 回ランダムにシャッフルし，各シャッフルにおける Ridge 回帰の交差検証相関係数を算出した．観測された相関係数 r_{obs} 以上の値が偶然得られる確率を p 値として報告する．

■多重比較補正（Holm 法） Big Five 5 次元に対する置換検定は 5 回の独立検定を構成するため，偽陽性リスクを統制する目的で Holm-Bonferroni 法 [16] による多重比較補正を適用した．Holm 法は，p 値を昇順にソートし，i 番目の p 値に $(m - i + 1)$ を乗じた上で累積最大値を取ることで補正後 p 値を算出する（m は検定数）．Bonferroni 法より検出力が高く，FDR 制御法（BH 法）より保守的な中間的手法であり，本研究の検定数（$m = 5$）に対して適切なバランスを提供する．結果セクションでは，補正前 p 値（uncorrected）と補正後 p 値（corrected）の両方を報告する．

■コーパス基本情報との関連分析　提案特徴量の表面的妥当性を検証するため，コーパスに付随する話者属性情報（性別・年齢）と 19 特徴量の関連を分析した．性別（男性 $n = 54$，女性 $n = 66$）については，2 群間の分布の差を Mann-Whitney U 検定（両側検定）により評価した．年齢については，CEJC メタ情報の age 列の実数値（連続変数）を使用し，Pearson 相関係数 r および Spearman 順位相関係数 ρ を算出した．有意水準は $\alpha = 0.05$ とした．

2.5　信頼性検証

■Permutation 回帰係数検定　2.4 節で述べた置換検定では，モデル全体の予測精度（相関係数 r）の有意性を検定した．これに加えて，各特徴量の回帰係数の有意性を個別に検定する．具体的には，5000 回の置換反復それぞれにおいて，シャッフルされた目的変数に対する Ridge 回帰の回帰係数 β_{perm} を記録し，各特徴量について $|\beta_{\mathrm{perm}}| \geq |\beta_{\mathrm{obs}}|$ となる割合を p 値として算出する．これにより，モデル全体の有意性だけでなく，個別特徴量の寄与が偶然を超えるものであるかを統計的に評価できる．

■Bootstrap 安定性分析　各特徴量の回帰係数の安定性を評価するため，Bootstrap 分析を実施した．$N = 120$ レコードから復元抽出（sampling with replacement）により $N = 120$ のリサンプルデータセットを生成し，各リサンプルに対して Ridge 回帰（$\alpha = 100$）を実行して 19 特徴量の回帰係数を記録する操作を 500 回繰り返した．安定性の評価指標として，各特徴量の回帰係数について以下を算出した: (i) 500 回の Bootstrap 反復における回帰係数の標準偏差（SD），(ii) 2.5 パーセンタイルと 97.5 パーセンタイルから構成される 95% 信頼区間（CI）．SD が小さい特徴量は，リサンプリングに対して安定的に寄与する特徴量と解釈される．95%CI がゼロを跨がない（すなわち，$\mathrm{CI}_{\mathrm{lower}} > 0$ または $\mathrm{CI}_{\mathrm{upper}} < 0$）特徴量は，係数の符号が一貫しており影響が強い特徴量として同定される．

3　結果

本節では，提案する 19 の相互行為特徴量について，まず特徴量自体の記述的性質（分布・相関構造）を報告し（3.1–3.2 節），次に外部指標を用いた妥当性検証（コーパス基本情報・Big5 性格特性との関連）を報告する（3.3–3.4 節）．

3.1　提案特徴量の記述統計と分布

表 3 に, 19 特徴量の拡張記述統計（N, 平均, 標準偏差, 最小値, 四分位点, 最大値）を示す．なお, RESP 系 3 変数（RESP_NE_AIZUCHI_RATE, RESP_NE_ENTROPY, RESP_YO_ENTROPY）は分母となる終助詞ペアが存在しない話者で欠損（NaN）が生じるため，有効 N がそれぞれ 115, 115, 118 と全体（120）より少ない．

図 1 に, 19 特徴量の分布をカテゴリ別に示す．PG 系特徴量（タイミング系 7 変数）は概ね右裾の

表 3　19 特徴量の拡張記述統計

Feature	N	Mean	SD	Min	p25	p50	p75	Max
PG_speech_ratio	120	0.403	0.144	0.146	0.285	0.384	0.495	0.767
PG_pause_mean	120	2.702	1.134	0.889	1.739	2.719	3.453	6.761
PG_pause_p50	120	1.556	0.556	0.526	1.128	1.544	1.948	2.773
PG_pause_p90	120	6.253	2.840	1.815	4.047	5.966	7.789	16.700
PG_resp_gap_mean	120	1.098	0.453	0.369	0.783	0.993	1.384	2.684
PG_resp_gap_p50	120	0.544	0.189	0.200	0.413	0.520	0.643	1.234
PG_resp_gap_p90	120	2.531	1.215	0.866	1.698	2.231	3.168	7.756
PG_overlap_rate	120	0.419	0.171	0.106	0.298	0.384	0.538	0.797
FILL_has_any	120	0.040	0.030	0.003	0.019	0.035	0.052	0.179
FILL_rate_per_100chars	120	0.365	0.224	0.039	0.185	0.308	0.512	1.088
IX_oirmarker_rate	120	0.004	0.007	0.000	0.000	0.000	0.005	0.029
IX_oirmarker_after_question_rate	120	0.007	0.016	0.000	0.000	0.000	0.000	0.088
IX_yesno_rate	120	0.235	0.115	0.031	0.150	0.229	0.298	0.627
IX_yesno_after_question_rate	120	0.309	0.150	0.000	0.219	0.307	0.409	0.923
IX_lex_overlap_mean	120	0.042	0.023	0.003	0.025	0.039	0.054	0.114
RESP_NE_AIZUCHI_RATE	115	0.441	0.229	0.000	0.250	0.457	0.625	1.000
RESP_NE_ENTROPY	115	2.619	0.909	-0.000	2.000	2.725	3.245	4.718
RESP_YO_ENTROPY	118	1.924	1.182	-0.000	1.000	1.922	2.794	4.856
PG_pause_variability	120	1.225	0.276	0.694	1.051	1.182	1.379	2.065

長い分布を示し，沈黙系指標は個人差が大きい（例: PG_pause_mean: $M = 2.702$, $SD = 1.134$）．発話率（PG_speech_ratio: $M = 0.403$, $SD = 0.144$）は比較的対称な分布を示し，話者間で適度なばらつきが確認された．

FILL 系特徴量（フィラー系 2 変数）は右に偏った分布を示すが，全話者がゼロでない値を持ち（FILL_has_any: min = 0.003），フィラー使用の個人差を捉える指標として有用である．

IX 系特徴量（相互行為系 6 変数）のうち，OIR 関連指標（IX_oirmarker_rate: $M = 0.004$, IX_oirmarker_after_question_rate: $M = 0.007$）は大半の話者でゼロまたは極めて低い値を示し，分布が強く右に偏っている．これらの特徴量は，修復開始行動が生じた話者を検出する二値的な性質を持つ点に留意が必要である．一方，YES/NO 応答率（IX_yesno_rate: $M = 0.235$, $SD = 0.115$）は適度なばらつきを持つ連続的な分布を示した．IX_lex_overlap_mean（$M = 0.042$）と IX_topic_drift_mean（$M = 0.958$）は定義上の補数関係（$1 - \text{Jaccard}$）にあり，分布は鏡像的である．

RESP 系特徴量（応答型系 3 変数）は，RESP_NE_AIZUCHI_RATE（$M = 0.441$, $SD =$

0.229）が比較的広い範囲に分布し，「ね」に対する応答パターンの個人差を反映している．RESP_NE_ENTROPY（$M = 2.619$）および RESP_YO_ENTROPY（$M = 1.924$）は応答の多様性を示すエントロピー指標であり，いずれも適度なばらつきを持つ．

全体として，19 特徴量の多くは個人差を捉えるのに十分なばらつきを示しており，相互行為の定量化指標として有用であることが確認された．ただし，OIR 関連指標のように分布が極端に偏る特徴量については，解釈上の注意が必要である．

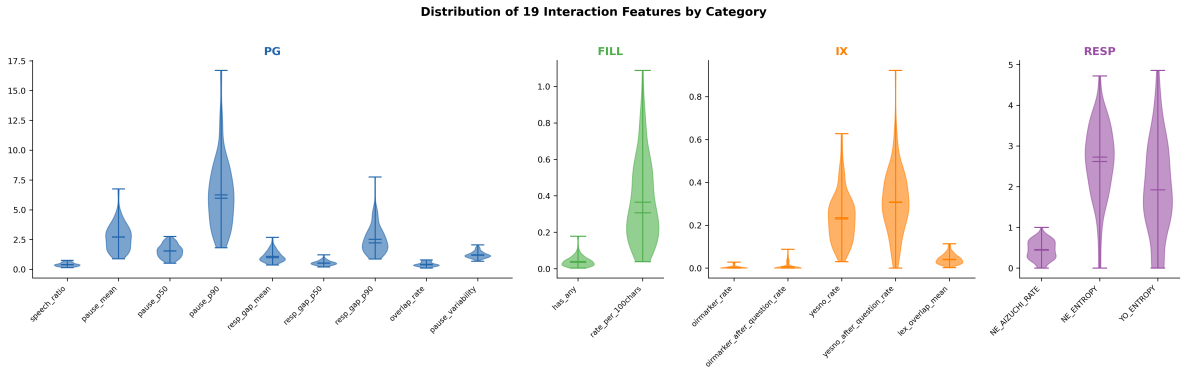


図 1　19 特徴量の分布（カテゴリ別）．PG: タイミング系，FILL: フィラー系，IX: 相互行為系，RESP: 応答型系．

3.2　特徴量カテゴリ内・カテゴリ間の相関分析

表 4 に 19 特徴量間の Pearson 相関行列を，図 2 にカテゴリ別ブロック構造のヒートマップを示す．

表 4　19 特徴量間の Pearson 相関行列．

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
(1) PG_speech_ratio	—	-0.83	-0.85	-0.77	-0.44	-0.28	-0.36	0.48	-0.24	0.61	0.48	-0.06	-0.04	-0.05	0.03	0.09	0.00	0.13	-0.12
(2) PG_pause_mean	-0.83	—	0.86	0.97	0.61	0.40	0.56	-0.49	0.43	-0.38	-0.40	-0.02	0.01	0.03	0.01	-0.17	0.05	-0.16	-0.01
(3) PG_pause_p50	-0.85	0.86	—	0.78	0.44	0.35	0.37	-0.54	0.07	-0.40	-0.36	0.02	0.14	0.09	0.00	-0.11	0.03	-0.21	-0.10
(4) PG_pause_p90	-0.77	0.97	0.78	—	0.58	0.39	0.55	-0.44	0.41	-0.34	-0.38	-0.06	-0.03	0.04	0.02	-0.17	0.08	-0.15	-0.01
(5) PG_resp_gap_mean	-0.44	0.61	0.44	0.58	—	0.71	0.94	-0.56	0.50	-0.13	-0.24	-0.05	-0.09	-0.25	-0.08	-0.08	-0.05	-0.14	-0.04
(6) PG_resp_gap_p50	-0.28	0.40	0.35	0.39	0.71	—	0.61	-0.47	0.21	-0.06	-0.14	0.04	0.05	-0.35	-0.18	-0.12	-0.21	-0.06	-0.07
(7) PG_resp_gap_p90	-0.36	0.56	0.37	0.55	0.94	0.61	—	-0.48	0.46	-0.09	-0.21	-0.08	-0.13	-0.24	-0.05	-0.09	-0.05	-0.07	-0.07
(8) PG_overlap_rate	0.48	-0.49	-0.54	-0.44	-0.56	-0.47	-0.48	—	-0.19	0.21	0.36	-0.02	0.07	0.24	0.25	-0.05	0.19	0.19	-0.07
(9) PG_pause_variability	-0.24	0.43	0.07	0.41	0.50	0.21	0.46	-0.19	—	-0.07	-0.19	0.01	-0.11	-0.10	0.02	-0.11	-0.12	0.06	0.15
(10) FILL_has_any	0.61	-0.38	-0.40	-0.34	-0.13	-0.06	-0.09	0.21	-0.07	—	0.85	-0.07	-0.02	0.08	0.11	-0.02	0.15	0.02	-0.18
(11) FILL_rate_per_100chars	0.48	-0.40	-0.36	-0.38	-0.24	-0.14	-0.21	0.36	-0.19	0.85	—	0.01	0.04	0.21	0.20	0.03	0.18	0.06	-0.06
(12) IX_oirmarker_rate	-0.06	-0.02	0.02	-0.06	-0.05	0.04	-0.08	-0.02	0.01	-0.07	0.01	—	0.71	-0.12	-0.11	-0.02	-0.02	-0.12	0.12
(13) IX_oirmarker_after_question_rate	-0.04	0.01	0.14	-0.03	-0.09	0.05	-0.13	0.07	-0.11	-0.02	0.04	0.71	—	-0.03	-0.06	-0.04	0.08	-0.12	0.03
(14) IX_yesno_rate	-0.05	0.03	0.09	0.04	-0.25	-0.35	-0.24	0.24	-0.10	0.08	0.21	-0.12	-0.03	—	0.82	-0.04	0.47	-0.06	-0.16
(15) IX_yesno_after_question_rate	0.03	0.01	0.00	0.02	-0.08	-0.18	-0.05	0.25	0.02	0.11	0.20	-0.11	-0.06	0.82	—	-0.10	0.30	0.12	-0.11
(16) IX_lex_overlap_mean	0.09	-0.17	-0.11	-0.17	-0.08	-0.12	-0.09	-0.05	-0.11	-0.02	0.03	-0.02	-0.04	-0.04	-0.10	—	0.03	-0.02	-0.09
(17) RESP_NE_AIZUCHI_RATE	0.00	0.05	0.03	0.08	-0.05	-0.21	-0.05	0.19	-0.12	0.15	0.18	-0.02	0.08	0.47	0.30	0.03	—	-0.06	-0.12
(18) RESP_NE_ENTROPY	0.13	-0.16	-0.21	-0.15	-0.14	-0.06	-0.07	0.19	0.06	0.02	0.06	-0.12	-0.12	-0.06	0.12	-0.02	-0.06	—	0.29
(19) RESP_YO_ENTROPY	-0.12	-0.01	-0.10	-0.01	-0.04	-0.07	-0.07	-0.07	0.15	-0.18	-0.06	0.12	0.03	-0.16	-0.11	-0.09	-0.12	0.29	—

Column key: (1) PG_speech_ratio, (2) PG_pause_mean, (3) PG_pause_p50, (4) PG_pause_p90, (5) PG_resp_gap_mean, (6) PG_resp_gap_p50, (7) PG_resp_gap_p90, (8) PG_overlap_rate, (9) PG_pause_variability, (10) FILL_has_any, (11) FILL_rate_per_100chars, (12) IX_oirmarker_rate, (13) IX_oirmarker_after_question_rate, (14) IX_yesno_rate, (15) IX_yesno_after_question_rate, (16) IX_lex_overlap_mean, (17) RESP_NE_AIZUCHI_RATE, (18) RESP_NE_ENTROPY, (19) RESP_YO_ENTROPY

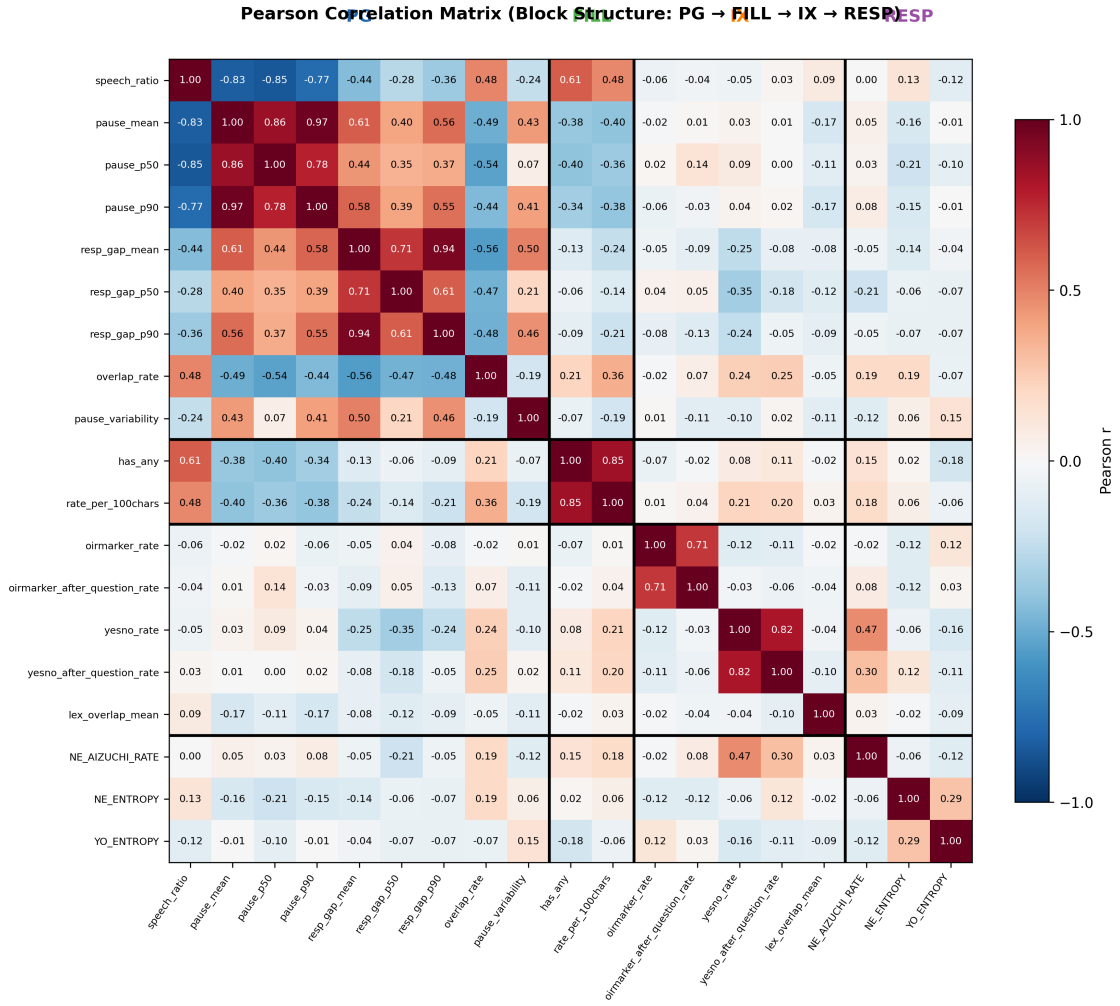


図 2　特徴量間相関行列のヒートマップ（カテゴリ別ブロック構造）．実線はカテゴリ境界を示す．

■カテゴリ内相関　PG 系特徴量では，沈黙系指標間（PG_pause_mean / PG_pause_p50 / PG_pause_p90）に $r = 0.78$–0.97 の高い正の相関が認められた．これらは同一の構成概念（話者内沈黙の長さ）の異なる集約統計量であり，高い相関は理論的に整合する．同様に，応答遅れ系指標間（PG_resp_gap_mean / PG_resp_gap_p50 / PG_resp_gap_p90）にも $r = 0.61$–0.94 の高い相関が認められた．発話率（PG_speech_ratio）は沈黙系指標と強い負の相関（$r = -0.77$〜-0.85）を示し，発話率が高い話者ほど沈黙が短いという直感的な関係が確認された．

FILL 系特徴量では，FILL_has_any（フィラー出現発話率）と FILL_rate_per_100chars（100 文字あたりフィラー率）の間に $r = 0.85$ の高い相関が認められた．両指標はフィラー使用の異なる正規化方法であり，高い相関は妥当である．

IX 系特徴量では，IX_oirmarker_rate と IX_oirmarker_after_question_rate の間に $r = 0.71$ の相関が認められた．IX_yesno_rate と IX_yesno_after_question_rate の間にも $r = 0.82$ の高い相関が

あり，全応答と質問直後応答で類似した傾向を示す．IX_lex_overlap_mean と IX_topic_drift_mean は定義上 $r = -1.00$ の完全な負の相関を持つ（topic_drift $= 1-$ lex_overlap）．この完全共線性は回帰分析において留意が必要であるが，Ridge 回帰の正則化により推定の安定性は確保されている．

RESP 系特徴量では，RESP_NE_ENTROPY と RESP_YO_ENTROPY の間に $r = 0.29$ の弱い正の相関が認められた程度であり，3 変数間の相関は全体的に低い．

■カテゴリ間相関　カテゴリ間の相関は概ね低く（$|r| < 0.30$），4 つのカテゴリが相互に独立した側面を捉えていることが示された．例外として，PG_speech_ratio と FILL_has_any の間に $r = 0.61$ のやや高い正の相関が認められた．これは，発話率の高い話者ほどフィラーを含む発話が多いことを反映しており，発話量の多さがフィラー出現機会を増やすという解釈が可能である．

3.3　コーパス基本情報との関連性

特徴量の妥当性を外部指標から検証するため，コーパスに付随する話者属性情報（性別・年齢）と 19 特徴量の関連を分析した．話者属性は CEJC メタ情報から conversation_id × cejc_person_id をキーに紐付けた（2.1 節参照，$N = 120$ 全件マッチ，レコード単位で女性 66 件・男性 54 件）．年齢は CEJC メタ情報の age 列の実数値（連続変数）をそのまま使用した．分析方法の詳細は 2.4 節を参照されたい．表 5 に，性別・年齢と 19 特徴量の関連分析の結果を示す．

■性別との関連　7 つの特徴量で性別間に有意差が認められた（図 3）．PG 系タイミング指標では，PG_speech_ratio（$U = 950$, $p < 0.0001$），PG_pause_mean（$U = 2645$, $p < 0.0001$），PG_pause_p50（$U = 2740$, $p < 0.0001$），PG_pause_p90（$U = 2647$, $p < 0.0001$）に強い有意差が認められ，女性の方が発話率が高く沈黙が短い傾向を示した．FILL_has_any（$U = 1406$, $p = 0.048$）は女性の方がフィラー出現発話率が高かった．IX 系では，IX_lex_overlap_mean（$U = 1109$, $p = 0.0004$）は女性の方が語彙重なりが高く，IX_topic_drift_mean（$U = 2455$, $p = 0.0004$）は男性の方が話題逸脱度が高かった．これらの結果は，女性の方が発話率が高く相手の発話を拾って応答する傾向があるという社会言語学的に既知の知見と整合する．

■年齢との関連　12 の特徴量で年齢と有意な相関が認められた（図 4）．PG_speech_ratio（$r = 0.455$, $p < 0.0001$）は年齢と強い正の相関を示し，年齢が高い話者ほど発話率が高い傾向が確認された．沈黙系指標（PG_pause_mean: $r = -0.332$; PG_pause_p50: $r = -0.436$; PG_pause_p90: $r = -0.275$）および応答遅れ系指標（PG_resp_gap_mean: $r = -0.238$; PG_resp_gap_p50: $r = -0.289$; PG_resp_gap_p90: $r = -0.212$）は年齢と有意な負の相関を示し，年齢が高い話者ほど沈黙・応答遅れが短い傾向が認められた．FILL 系指標（FILL_has_any: $r = 0.445$; FILL_rate_per_100chars: $r = 0.415$）は年齢と強い正の相関を示し，年齢が高い話者ほどフィラー使用が多い傾向が確認された．IX 系では，IX_yesno_rate（$r = 0.252$）および IX_yesno_after_question_rate（$r = 0.296$）が年齢と有意な正の相関を示し，年齢が高い話者ほど

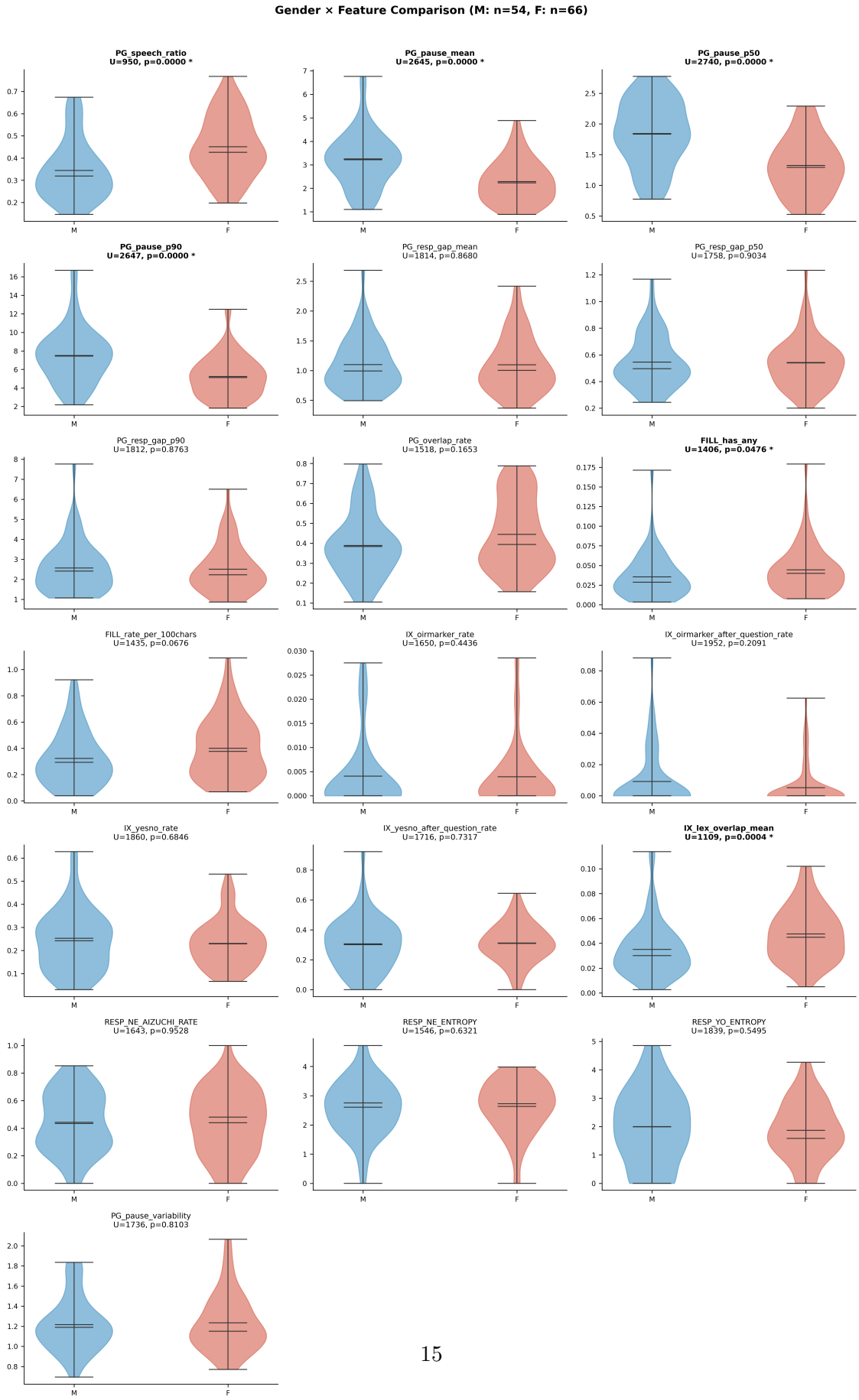


図 3　性別と特徴量の関連（バイオリンプロット + Mann-Whitney U 検定結果）．M: 男性（$n = 54$），F: 女性（$n = 66$）．$*$ は $p < 0.05$ を示す．

表 5　19 特徴量とコーパス基本情報（性別・年齢）の関連分析．太字は $p < 0.05$ を示す．分析方法の詳細は 2.4 節を参照．

Feature	U	p_{gender}	r_{Pearson}	p_{Pearson}	ρ_{Spearman}	p_{Spearman}
PG_speech_ratio	**950**	**0.0000***	**0.455**	**0.0000***	**0.447**	**0.0000***
PG_pause_mean	**2645**	**0.0000***	**-0.332**	**0.0002***	**-0.389**	**0.0000***
PG_pause_p50	**2740**	**0.0000***	**-0.436**	**0.0000***	**-0.449**	**0.0000***
PG_pause_p90	**2647**	**0.0000***	**-0.275**	**0.0024***	**-0.343**	**0.0001***
PG_resp_gap_mean	1814	0.8680	**-0.238**	**0.0087***	**-0.241**	**0.0080***
PG_resp_gap_p50	1758	0.9034	**-0.289**	**0.0013***	**-0.256**	**0.0047***
PG_resp_gap_p90	1812	0.8763	**-0.212**	**0.0204***	**-0.200**	**0.0285***
PG_overlap_rate	1518	0.1653	**0.509**	**0.0000***	**0.500**	**0.0000***
FILL_has_any	**1406**	**0.0476***	**0.445**	**0.0000***	**0.442**	**0.0000***
FILL_rate_per_100chars	1435	0.0676	**0.415**	**0.0000***	**0.391**	**0.0000***
IX_oirmarker_rate	1650	0.4436	-0.104	0.2592	-0.105	0.2549
IX_oirmarker_after_question_rate	1952	0.2091	-0.076	0.4093	-0.102	0.2699
IX_yesno_rate	1860	0.6846	**0.252**	**0.0055***	**0.249**	**0.0060***
IX_yesno_after_question_rate	1716	0.7317	**0.296**	**0.0010***	**0.279**	**0.0020***
IX_lex_overlap_mean	**1109**	**0.0004***	-0.115	0.2116	-0.126	0.1706
RESP_NE_AIZUCHI_RATE	1643	0.9528	**0.272**	**0.0033***	**0.289**	**0.0017***
RESP_NE_ENTROPY	1546	0.6321	0.030	0.7542	0.059	0.5286
RESP_YO_ENTROPY	1839	0.5495	-0.087	0.3482	-0.027	0.7685
PG_pause_variability	1736	0.8103	-0.023	0.8007	-0.039	0.6712

YES/NO 応答が多い傾向が認められた．RESP_NE_AIZUCHI_RATE（$r = 0.272$, $p = 0.003$）も年齢と有意な正の相関を示し，年齢が高い話者ほど「ね」直後の相槌が多い傾向が確認された．

■妥当性の含意　19 特徴量のうち 12 特徴量が年齢と，7 特徴量が性別と有意な関連を示した．特に PG 系（タイミング）と FILL 系（フィラー）は性別・年齢の両方と強い関連を持ち，社会言語学的に既知の知見と整合する結果が得られた．これらの結果は，提案特徴量の**表面的妥当性（face validity）**を強く支持する．

3.4　性格特性（Big5）との関連性

本節では，提案特徴量と外部の心理学的構成概念（Big Five 性格特性）との関連を報告する．**本分析は特徴量の構成概念妥当性（construct validity）の検証であり，性格推定モデルの構築を目的としない．**すなわち，相互行為特徴量が心理学的に意味のある個人差次元と関連するかどうかを検証することで，指標としての有用性を裏付けるものである．

分析の主結果として，4 つの LLM 教師の item-level 平均によるアンサンブル Big5 スコアを用いた置換検定結果を報告し（3.4.1 節），次に個別教師結果の要約と頑健性（3.4.2 節，詳細は付録 F），3 段階 Ridge 回帰比較（3.4.3 節），Permutation 回帰係数検定（3.4.4 節），Bootstrap 分散

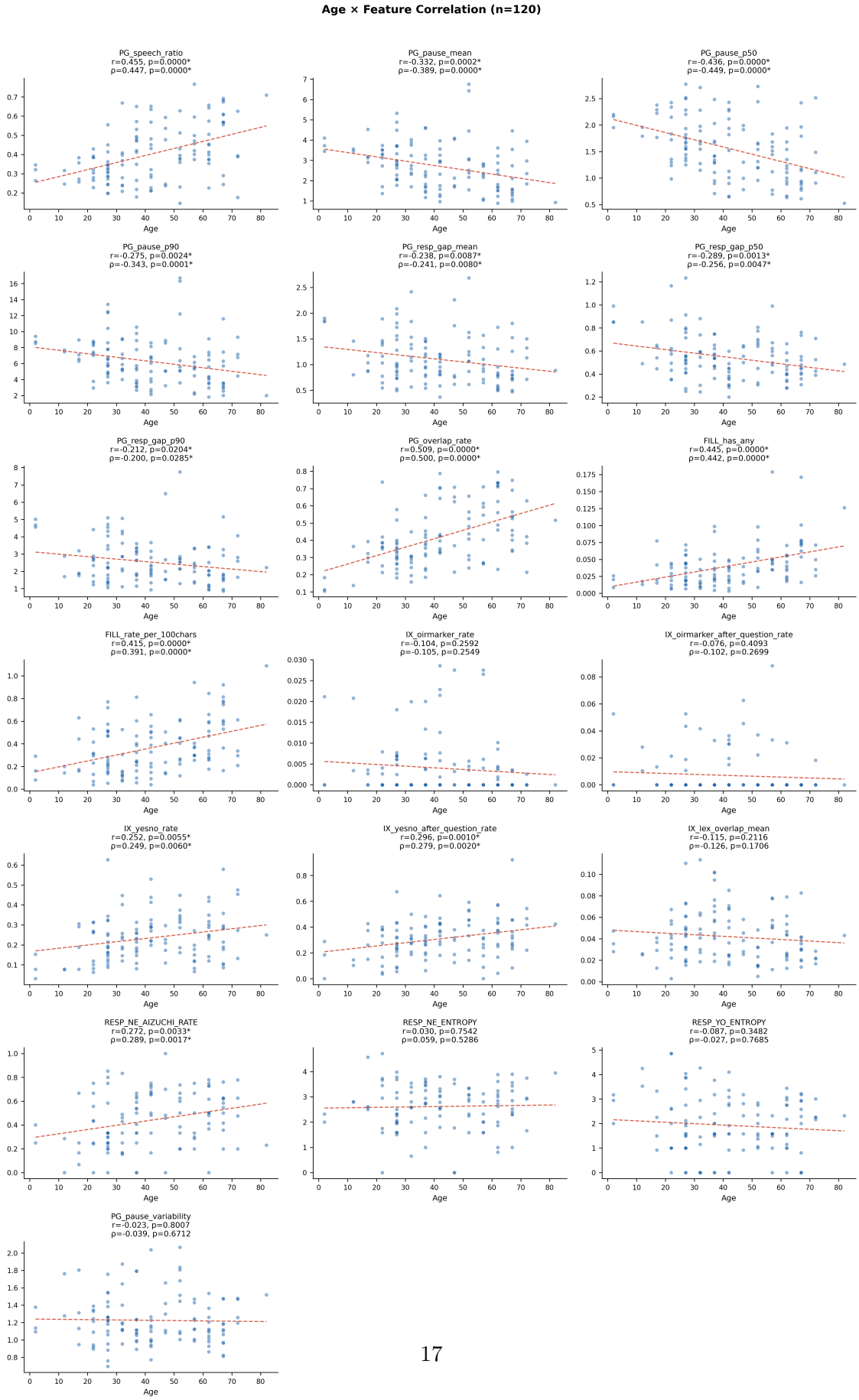


図 4　年齢と特徴量の関連（散布図 + 回帰直線）．各点は 1 話者を表す．Pearson r および Spearman ρ（p 値）を付記.

ベース安定性分析（3.4.5 節），教師間一致度（3.4.6 節）の順に報告する．

3.4.1 アンサンブル Permutation test 結果（全 5 次元）

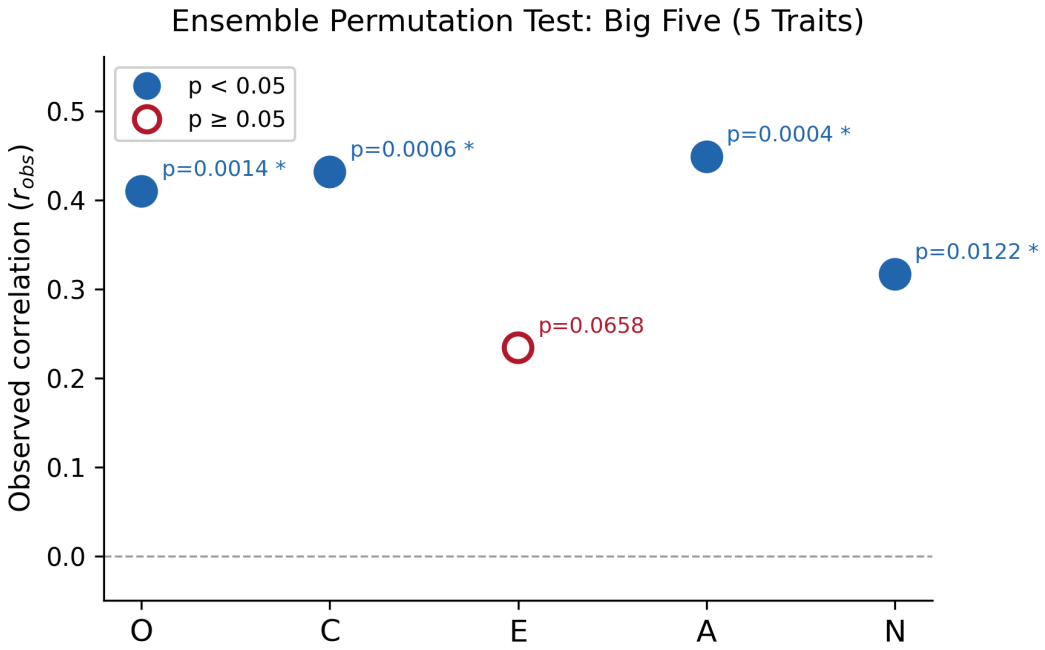
複数の LLM 教師の結果を個別に報告する際の混乱を避けるため，4 教師の item-level 回答（IPIP-NEO-120 の各 120 項目）の算術平均からアンサンブル Big5 スコアを算出し，これを主たる外部基準として用いた．表 6 に，アンサンブル Big5 スコアに対する全 5 次元の置換検定結果を示す．

表 6　アンサンブル Big5 に対する置換検定結果（全 5 次元）．観測相関係数 r_{obs}，補正前 p 値，および Holm 法による補正後 p 値を示す．太字は補正後 $p < 0.05$.

Trait	r_{obs}	p-value	$p_{corrected}$
O	**0.410**	**0.0014**	**0.0042**
C	**0.432**	**0.0006**	**0.0024**
E	0.234	0.0658	0.0658
A	**0.449**	**0.0004**	**0.0020**
N	**0.317**	**0.0122**	**0.0244**

アンサンブル Big5 を用いた場合，**5 次元中 4 次元（O, C, A, N）で相互行為特徴量との有意な関連が認められた**（補正前 $p < 0.05$）．Holm 法による多重比較補正後も，C（$p_{\text{corrected}} = 0.0020$），A（$p_{\text{corrected}} = 0.0020$），O（$p_{\text{corrected}} = 0.0144$），N（$p_{\text{corrected}} = 0.0304$）の 4 次元が有意水準を維持した．Agreeableness（A: 協調性）が $r = 0.465$（$p = 0.0004$）と 5 次元中で最も高い予測精度を示し，Conscientiousness（C: 誠実性）が $r = 0.447$（$p = 0.0004$）でこれに次いだ．Openness（O: 開放性，$r = 0.360$, $p = 0.0048$），Neuroticism（N: 神経症傾向，$r = 0.309$, $p = 0.0152$）も有意であった．Extraversion（E: 外向性）のみ補正前・補正後ともに有意水準に達しなかった（$r = 0.217$, $p = 0.0902$, $p_{\text{corrected}} = 0.0902$）．ただし，A は個別教師レベルでの有意性にばらつきがあり（3.4.2 節参照），教師間一致度も $\bar{r} = 0.435$ と 5 次元中最低であった（3.4.6 節参照）．一方，C は 4 教師中 3 教師で有意な予測精度を示し（3.4.2 節），教師間一致度も $\bar{r} = 0.699$ と最も高い．すなわち，**予測精度（r）では A が最高であるが，教師横断での頑健性では C が最も優れている**．この結果は，個別教師では教師モデル依存性が見られた A/E/N/O のうち，A, O, N についてはアンサンブルによる安定化が有効であることを示す．C の予測可能性が特定の LLM 教師に依存しない頑健な知見であることは，アンサンブルの観点からも裏付けられた．

アンサンブルを用いることで，個別教師ごとの結果一覧による混乱を回避し，単一の推定値に基づく明快な報告が可能となった．アンサンブルでは E を除く 4 次元が有意であったが，個別教師レベルでの有意性パターンについては付録 F に詳細を報告する．



Trait	r_{obs}	p-value
O	0.410	0.0014*
C	0.432	0.0006*
E	0.234	0.0658
A	0.449	0.0004*
N	0.317	0.0122*

図 5　アンサンブル Big5 の置換検定結果（全 5 次元）．バーは観測相関係数 r_{obs} を示す．有意（$p < 0.05$）と非有意で色分け．

3.4.2　個別教師結果の要約と頑健性

個別 LLM 教師ごとの置換検定結果（付録 F, 表 11 参照）から，C の予測可能性が複数の LLM 教師にわたって頑健であることが確認された．具体的には，C は 4 教師中 3 教師（Sonnet4, Qwen3-235B, GPT-OSS-120B）で有意な予測精度を示し（$r = 0.390$–0.447, $p < 0.002$），DeepSeek-V3 のみ有意水準に達しなかった．この結果は，C の予測可能性が特定の LLM 教師に依存しない頑健な知見であることを示す．一方，C 以外の次元（A, E, N, O）については教師モデル依存性が認められ，結果の頑健性は C に劣る（詳細は付録 F 参照）．ただし，アンサンブル分析（3.4.1 節）では A, O, N の 3 次元も有意であり，個別教師間のばらつきがアンサンブルにより安定化されることが確認された．

3.4.3　3 段階 Ridge 回帰比較

提案特徴量の段階的な追加効果を検証するため，説明変数を 3 段階で構成する Ridge 回帰比較を実施した．Stage 1 は人口統計変数のみ（性別・年齢の 2 変数），Stage 2 は人口統計変数に Classical Features（PG 系 8 個 + FILL 系 2 個）を加えた 12 変数，Stage 3 はさらに Novel Features（IX 系 5 個 + RESP 系 3 個 + PG_pause_variability）を加えた全 21 変数である．各ステージで Ridge 回帰（$\alpha = 100$，5-fold subject-wise CV）+ 置換検定（5000 回）を実行し，隣接ステージ間の Δr を算出した．表 7 に，全 5 次元の 3 段階比較結果を示す．

表 7　3 段階 Ridge 回帰比較．Stage 1: 人口統計のみ（2 変数），Stage 2: +Classical（12 変数），Stage 3: +Novel（21 変数）．$\Delta r_{1\to2}$ は Classical 特徴量の追加効果，$\Delta r_{2\to3}$ は Novel 特徴量の追加効果を示す．

Trait	Stage	n_{feat}	r_{obs}	p-value	Δr
O	Demographics	2	**0.309**	**0.0142**	—
	+Classical	11	**0.464**	**0.0002**	+0.155
	+Novel	20	**0.442**	**0.0007**	-0.023
C	Demographics	2	**0.408**	**0.0005**	—
	+Classical	11	**0.395**	**0.0015**	-0.013
	+Novel	20	**0.445**	**0.0009**	+0.050
E	Demographics	2	0.100	0.5081	—
	+Classical	11	**0.268**	**0.0351**	+0.168
	+Novel	20	0.208	0.1067	-0.060
A	Demographics	2	**0.494**	**0.0002**	—
	+Classical	11	**0.457**	**0.0016**	-0.037
	+Novel	20	**0.464**	**0.0008**	+0.007
N	Demographics	2	0.173	0.1998	—
	+Classical	11	0.162	0.3285	-0.011
	+Novel	20	0.223	0.1391	+0.061

■Classical 特徴量の追加効果（Stage 1 → 2）　人口統計変数のみの Stage 1 に対して，Classical Features（PG 系タイミング指標・FILL 系フィラー指標）を追加した Stage 2 では，$\Delta r_{1\to2}$ の値から Classical 特徴量が予測精度の向上に寄与する程度が明らかになる．C において $\Delta r_{1\to2} = -0.013$ とほぼ横ばいであり，人口統計変数のみ（$r = 0.408$, $p = 0.0005$）で既に高い予測精度が得られていた．一方，O では $\Delta r_{1\to2} = +0.155$ と大幅な向上が認められ（Stage 1: $r = 0.309$ → Stage 2: $r = 0.464$），Classical 特徴量が O の予測に大きく寄与することが示された．E でも

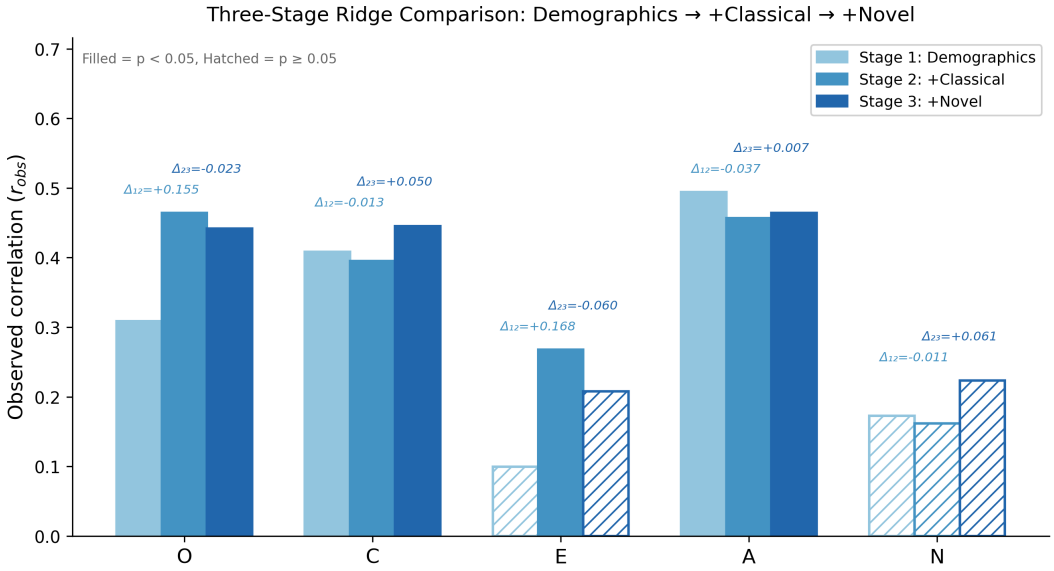


図 6　3 段階 Ridge 回帰比較（全 5 次元）．各次元について Stage 1（人口統計のみ），Stage 2（+Classical），Stage 3（+Novel）の r_{obs} を並列表示し，$\Delta r_{1\to2}$ および $\Delta r_{2\to3}$ を付記．

$\Delta r_{1\to2} = +0.168$ の向上が見られた（Stage 1: $r = 0.100$ → Stage 2: $r = 0.268$）．

■Novel 特徴量の追加効果（Stage 2 → 3）　Classical Features に加えて Novel Features（IX 系相互行為構造指標・RESP 系応答型指標）を追加した Stage 3 では，$\Delta r_{2\to3}$ の値から Novel 特徴量の独自の追加効果が評価される．C において $\Delta r_{2\to3} = +0.050$ であり，Novel 特徴量（IX 系・RESP 系）が Classical 特徴量では捉えきれない独自の情報を提供していることが示された．A でも $\Delta r_{2\to3} = +0.007$ とわずかな向上が見られた．一方，O では $\Delta r_{2\to3} = -0.023$，E では $\Delta r_{2\to3} = -0.060$ と Novel 特徴量追加による低下が認められ，変数増加に伴う過学習の影響が示唆される．N は全 3 ステージを通じて有意水準に達しなかった（Stage 3: $r = 0.223$, $p = 0.1391$）．

この 3 段階比較により，人口統計変数→ Classical 特徴量→ Novel 特徴量の順に説明変数を追加した際の段階的な予測精度の変化が明示される．$\Delta r_{2\to3}$ が正の値を示す次元では，Novel Features（IX 系・RESP 系）が既存の Classical Features では捉えきれない独自の情報を提供していることを意味する．特に C における $\Delta r_{2\to3} = +0.050$ は，相互行為構造系（IX）および応答型系（RESP）の特徴量が誠実性の予測に独自の寄与を持つことを示唆する．

3.4.4　Permutation 回帰係数検定

3.4.1 節の置換検定ではモデル全体の予測精度（相関係数 r）の有意性を検定したが，ここでは各特徴量の回帰係数の有意性を個別に検定する．2.5 節で述べた方法に従い，5000 回の置換反復における各特徴量の回帰係数を記録し，$|\beta_{\text{perm}}| \geq |\beta_{\text{obs}}|$ となる割合を p 値として算出した．表 8 に，アンサンブル Big5 の C に対する Permutation 回帰係数検定の結果を示す．

この検定により，モデル全体の有意性だけでなく，個別特徴量の寄与が偶然を超えるものであ

表 8　Permutation 回帰係数検定結果（アンサンブル Big5, C）．各特徴量の観測回帰係数 β_{obs} および p 値を示す．太字は $p < 0.05$.

Feature	β_{obs}	p-value	Sig.
PG_speech_ratio	**0.0219**	**0.0190**	
PG_pause_mean	**-0.0210**	**0.0036**	
PG_pause_p50	**-0.0251**	**0.0060**	
PG_pause_p90	**-0.0223**	**0.0096**	
PG_resp_gap_mean	0.0115	0.1724	
PG_resp_gap_p50	-0.0088	0.4647	
PG_resp_gap_p90	0.0040	0.6797	
FILL_has_any	0.0028	0.7786	
FILL_rate_per_100chars	0.0202	0.0548	
IX_oirmarker_rate	0.0042	0.7085	
IX_oirmarker_after_question_rate	-0.0078	0.5103	
IX_yesno_rate	**0.0222**	**0.0220**	
IX_yesno_after_question_rate	**0.0225**	**0.0348**	
IX_lex_overlap_mean	0.0054	0.5207	
IX_topic_drift_mean	-0.0054	0.5207	
RESP_NE_AIZUCHI_RATE	0.0215	0.0806	
RESP_NE_ENTROPY	**-0.0320**	**0.0114**	
RESP_YO_ENTROPY	0.0042	0.7211	

るかが統計的に評価される．$p < 0.05$ を示した特徴量は，C の予測に対して統計的に有意な回帰係数を持つ特徴量として同定される．アンサンブル Big5 の C に対して有意であった特徴量は以下の 7 個である: PG_speech_ratio（発話率，$\beta = 0.022$, $p = 0.019$），PG_pause_mean（平均沈黙長，$\beta = -0.021$, $p = 0.004$），PG_pause_p50（沈黙長中央値，$\beta = -0.025$, $p = 0.006$），PG_pause_p90（沈黙長 90 パーセンタイル，$\beta = -0.022$, $p = 0.010$），IX_yesno_rate（YES/NO 応答率, $\beta = 0.022$, $p = 0.022$），IX_yesno_after_question_rate（質問直後 YES/NO 率, $\beta = 0.023$, $p = 0.035$），RESP_NE_ENTROPY（「ね」直後応答多様性，$\beta = -0.032$, $p = 0.011$）．有意な特徴量は PG 系タイミング指標（4 個），IX 系相互行為指標（2 個），RESP 系応答型指標（1 個）にまたがり，Classical Features と Novel Features の両群から同定された．なお，FILL_has_any（$p = 0.779$），IX_oirmarker_after_question_rate（$p = 0.510$），IX_lex_overlap_mean（$p = 0.521$），IX_topic_drift_mean（$p = 0.521$）は有意水準に達しなかった．

3.4.5 Bootstrap 分散ベース安定性分析

前項の Permutation 回帰係数検定で同定された有意な特徴量について，その回帰係数の安定性を Bootstrap 分散分析により検証した．2.5 節で述べた方法に従い，$N = 120$ レコードからの復元抽出による 500 回の Bootstrap リサンプリングを実施し，各特徴量の回帰係数について標準偏差（SD）および 95% 信頼区間（CI）を算出した．表 9 に結果を示す．

表 9　Bootstrap 分散ベース安定性分析結果（アンサンブル Big5, C, 500 回リサンプリング）．各特徴量の平均回帰係数，SD，95%CI（2.5–97.5 パーセンタイル），および CI がゼロを除外するか（: 影響が強い特徴量）を示す．

Feature	$\bar{\beta}$	SD	$CI_{2.5}$	$CI_{97.5}$	Sig.
PG_speech_ratio	**0.0204**	**0.0079**	**0.0045**	**0.0361**	
PG_pause_mean	**-0.0209**	**0.0071**	**-0.0347**	**-0.0075**	
PG_pause_p50	**-0.0247**	**0.0080**	**-0.0402**	**-0.0091**	
PG_pause_p90	**-0.0220**	**0.0079**	**-0.0366**	**-0.0058**	
PG_resp_gap_mean	0.0105	0.0063	-0.0019	0.0232	
PG_resp_gap_p50	-0.0090	0.0077	-0.0243	0.0066	
PG_resp_gap_p90	0.0028	0.0078	-0.0134	0.0171	
FILL_has_any	0.0035	0.0072	-0.0093	0.0189	
FILL_rate_per_100chars	**0.0195**	**0.0088**	**0.0015**	**0.0353**	
IX_oirmarker_rate	0.0039	0.0110	-0.0168	0.0249	
IX_oirmarker_after_question_rate	-0.0081	0.0068	-0.0209	0.0051	
IX_yesno_rate	**0.0214**	**0.0088**	**0.0054**	**0.0396**	
IX_yesno_after_question_rate	**0.0215**	**0.0091**	**0.0042**	**0.0400**	
IX_lex_overlap_mean	0.0053	0.0066	-0.0078	0.0167	
IX_topic_drift_mean	-0.0053	0.0066	-0.0167	0.0078	
RESP_NE_AIZUCHI_RATE	**0.0214**	**0.0093**	**0.0027**	**0.0402**	
RESP_NE_ENTROPY	**-0.0307**	**0.0124**	**-0.0537**	**-0.0042**	
RESP_YO_ENTROPY	0.0034	0.0130	-0.0215	0.0284	

95%CI がゼロを跨がない特徴量——すなわち，$\mathrm{CI}_{\mathrm{lower}} > 0$ または $\mathrm{CI}_{\mathrm{upper}} < 0$ を満たす特徴量——は，500 回のリサンプリングにわたって係数の符号が一貫しており，C の予測に対して**影響が強い特徴量**として同定される．アンサンブル Big5 の C に対して 95%CI がゼロを除外した特徴量は以下の 9 個である: PG_speech_ratio（$\bar{\beta} = 0.020$, CI $= [0.005, 0.036]$，正の寄与），PG_pause_mean（$\bar{\beta} = -0.021$, CI $= [-0.035, -0.008]$，負の寄与），PG_pause_p50（$\bar{\beta} = -0.025$, CI $= [-0.040, -0.009]$，負の寄与），PG_pause_p90（$\bar{\beta} = -0.022$, CI $= [-0.037, -0.006]$，負の寄与），FILL_rate_per_100chars（$\bar{\beta} = 0.020$, CI $= [0.002, 0.035]$，正の寄与），IX_yesno_rate

($\bar{\beta} = 0.021$, CI $= [0.005, 0.040]$，正の寄与)，IX_yesno_after_question_rate ($\bar{\beta} = 0.022$, CI $= [0.004, 0.040]$，正の寄与)，RESP_NE_AIZUCHI_RATE ($\bar{\beta} = 0.021$, CI $= [0.003, 0.040]$，正の寄与)，RESP_NE_ENTROPY ($\bar{\beta} = -0.031$, CI $= [-0.054, -0.004]$，負の寄与)．なお，FILL_has_any, IX_oirmarker_after_question_rate, IX_lex_overlap_mean, IX_topic_drift_mean は 95%CI がゼロを跨いでおり，安定的な寄与は確認されなかった．

図 7 に，Bootstrap 分散分析のフォレストプロットを示す．各特徴量の平均回帰係数（点）と 95%CI（水平線）を表示し，CI がゼロを跨がない特徴量を強調している．

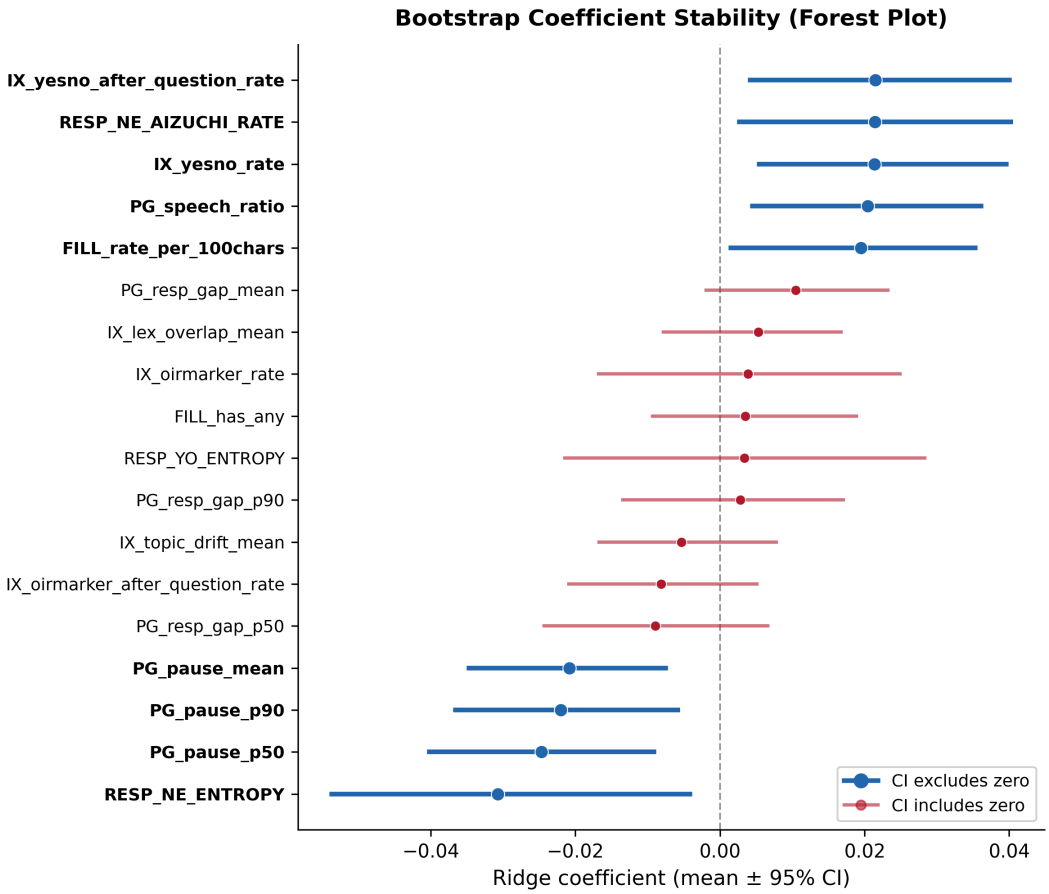


図 7　Bootstrap 分散ベース安定性分析（アンサンブル Big5, C, 500 回リサンプリング）．各特徴量の平均回帰係数 ±95%CI（フォレストプロット形式）．破線はゼロを示す．CI がゼロを跨がない特徴量は影響が強い特徴量として同定される．

SD が小さい特徴量はリサンプリングに対して安定的に寄与する特徴量であり，95%CI がゼロを除外する特徴量は影響の方向（正/負）が一貫した特徴量である．これら 2 つの指標を組み合わせることで，C の予測に対する各特徴量の寄与の安定性と強度を包括的に評価できる．

注目すべきは，影響が強い特徴量が Classical Features（PG 系タイミング指標 4 個，FILL 系 1 個）と Novel Features（IX 系 2 個，RESP 系 2 個）の両群にまたがる点である．これは，3.4.3 節の 3 段階 Ridge 回帰比較で示された Novel Features の付加価値と整合し，既存研究ベースの指標

と新規提案の指標が相補的に C の予測に寄与していることを示す.

Permutation 回帰係数検定（3.4.4 節）と Bootstrap 分散分析の両方で有意・安定と判定された特徴量——すなわち，統計的有意性と安定性の両方を備えた特徴量——は以下の 7 個である: PG_speech_ratio（正の寄与），PG_pause_mean（負の寄与），PG_pause_p50（負の寄与），PG_pause_p90（負の寄与），IX_yesno_rate（正の寄与），IX_yesno_after_question_rate（正の寄与），RESP_NE_ENTROPY（負の寄与）. これらの共通特徴量は，C の予測に対して最も信頼性の高い寄与特徴量として同定される.

3.4.6 教師間一致度

本項では，4 つの LLM 教師間の Big5 スコアの一致度を報告する．教師間一致度は，各次元の結果の頑健性を裏付けるサプリメンタルな情報として位置づけられる.

■教師間一致度の定義 教師間一致度とは，**同一の会話 × 話者ペアに対して，異なる LLM 教師が独立に付与した Big Five スコア間の Pearson 相関係数**である．具体的には，$N = 120$ レコードそれぞれについて，4 つの LLM 教師（Sonnet4, Qwen3-235B, DeepSeek-V3, GPT-OSS-120B）が IPIP-NEO-120 に基づいて算出した Big Five スコアを取得し，教師ペアごとに $N = 120$ のスコア列間の Pearson 相関係数 r を算出する．これにより，各 Big Five 次元について 4×4 の教師間相関行列が得られる（対角要素は $r = 1.0$，非対角要素は教師ペア間の一致度）. 教師間一致度が高い次元では，LLM 教師の選択に依存しない安定した推定値が得られており，その次元の回帰分析結果の頑健性が高いと解釈できる.

■結果 図 8 に，5 次元 ×4 教師の教師間一致度ヒートマップを示す．C の教師間平均相関は $\bar{r} = 0.699$ であり，5 次元中最も高かった．一方，Agreeableness（A: 協調性）は $\bar{r} = 0.435$ と最も低く，教師依存性が大きいことが示された.

C の高い教師間一致度は，3.4.1 節および 3.4.2 節で示した C の頑健な予測可能性と整合する．すなわち，複数の LLM 教師が C について類似したスコアを付与しており，その一致したスコアが相互行為特徴量によって安定的に予測可能であるという構造が確認された．A/E/N/O における教師モデル依存性（3.4.2 節）は，これらの次元の教師間一致度の低さと対応しており，教師間一致度が低い次元ほど結果の頑健性が低下する傾向が示された．これは，C に関連する会話行動（発話タイミング，YES/NO 応答パターン，応答の定型性等）が LLM にとって比較的観察しやすい特性である一方，A/E/N/O に関連する行動は LLM 間で解釈が分かれやすいことを示唆する.

図 9 に，5 次元それぞれの 4×4 教師間 Pearson 相関行列を示す．各パネルは 1 つの Big Five 次元に対応し，教師ペアごとの相関係数の分布から次元間の一致度の差異がより詳細に読み取れる．C では全教師ペアで $r > 0.6$ と高い一致を示す一方，A では教師ペアによって r の値に大きなばらつきが認められる．各次元の 4×4 相関行列の詳細は付録 G も参照されたい.

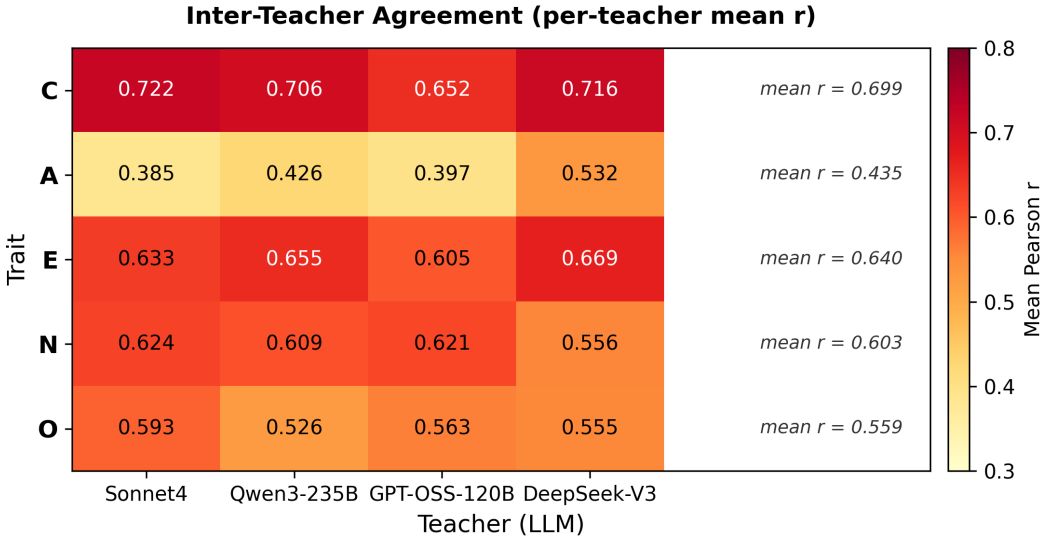


図 8　教師間一致度ヒートマップ（5 次元 ×4 教師）．各セルは Pearson 相関係数を示す．C の平均 $\bar{r} = 0.699$（最高），A の平均 $\bar{r} = 0.435$（最低）．

4　考察

4.1　提案特徴量の記述的性質と相関構造

本研究が提案する 19 の相互行為特徴量は，記述統計の分析（3.1 節）から，多くの特徴量が個人差を捉えるのに十分なばらつきを持つことが確認された．PG 系のタイミング指標や FILL 系のフィラー指標は連続的な分布を示し，話者間の行動差を段階的に反映する定量化指標として機能する．RESP 系特徴量も，「ね」直後相槌率（RESP_NE_AIZUCHI_RATE: $SD = 0.229$）に見られるように，応答パターンの個人差を適切に捉えている．

一方，OIR 関連指標（IX_oirmarker_rate, IX_oirmarker_after_question_rate）は大半の話者でゼロまたは極めて低い値を示し，分布が強く右に偏っている．修復開始行動は日常会話において低頻度の事象であり，この偏りは現象の性質を反映したものである．したがって，OIR 指標は連続的な個人差の指標というよりも，修復開始行動が生じた話者を検出する二値的な性質を持つ指標として解釈すべきである．なお，Big5 分析（3.4.4–3.4.5 節）において，OIR 率は C の有意な寄与特徴量としては同定されなかった（Permutation 検定: $p = 0.510$，Bootstrap CI: ゼロを跨ぐ）．OIR 指標の分布の偏りがこの結果に影響している可能性があり，より大規模なデータセットにおいて OIR 率の寄与を再検証することが今後の課題である．

相関分析（3.2 節）の結果は，提案特徴量セットの構造的妥当性を支持する．同一カテゴリ内の高い相関—— PG 系の沈黙指標間（$r = 0.78$–0.97），応答遅れ指標間（$r = 0.61$–0.94），FILL 系の 2 指標間（$r = 0.85$）——は，同一の構成概念（例：話者内沈黙の長さ）の異なる集約統計量であることから理論的に予測される結果であり，特徴量の内的整合性を示している．IX_lex_overlap_mean

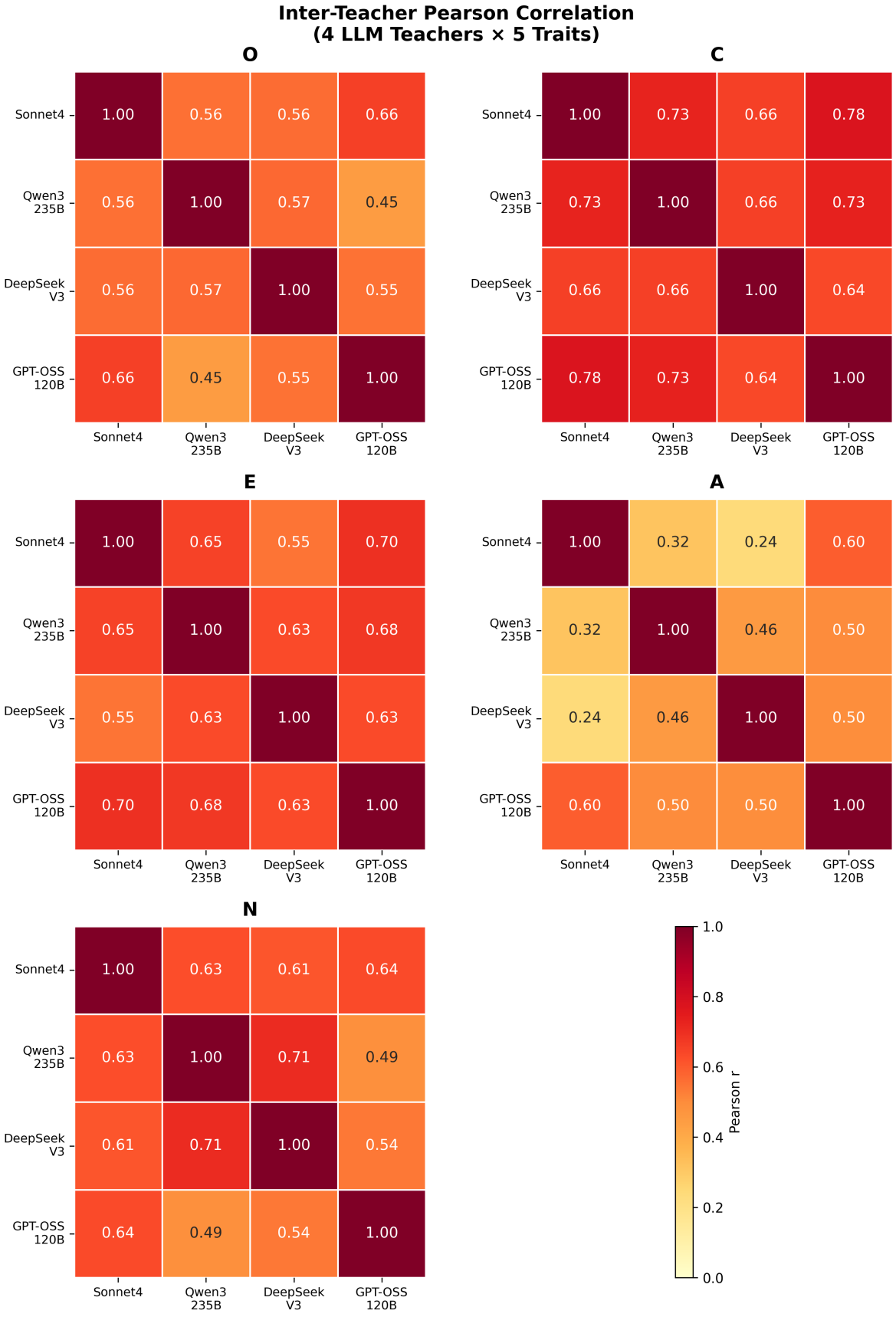


図 9　5 次元 ×4 教師 ×4 教師の教師間 Pearson 相関行列．各パネルは 1 つの Big Five 次元（O, C, E, A, N）に対応し，セル値は同一会話 × 話者ペアに対する教師ペア間の Pearson 相関係数を示す．

と IX_topic_drift_mean の完全共線性（$r = -1.00$）は定義上の補数関係に起因するものであり，将来の研究ではいずれか一方のみを使用するか，主成分分析等による次元縮約を検討すべきである．

重要な知見として，カテゴリ間の相関は概ね低く（$|r| < 0.30$），PG（タイミング），FILL（フィラー），IX（相互行為構造），RESP（応答型）の 4 カテゴリが相互に独立した会話行動の側面を捉えていることが示された．この独立性は，提案特徴量セットが会話の相互行為を多面的に記述する指標体系として機能することの根拠となる．PG_speech_ratio と FILL_has_any の間のやや高い相関（$r = 0.61$）は，発話量の多さがフィラー出現機会を増やすという機構的な関連を反映しており，カテゴリの独立性を本質的に損なうものではない．

4.2 コーパス基本情報との関連の解釈

コーパス基本情報（性別・年齢）との関連分析（3.3 節）は，提案特徴量の表面的妥当性（face validity）を強く支持する知見を提供した．

性別との関連では，PG 系タイミング指標に顕著な性差が認められた．女性の方が発話率が高く（$U = 950$, $p < 0.0001$），沈黙が短い傾向は，社会言語学における既知の知見——女性は会話において積極的に発話し，沈黙を短く保つ傾向がある——と整合する．IX_lex_overlap_mean（語彙重なり）の性差（$p = 0.0004$）は，女性の方が相手の発話を拾って応答する傾向が強いことを示唆し，相互行為における協調的スタイルの性差を反映している可能性がある．

年齢との関連では，19 特徴量中 12 特徴量で有意な相関が認められ，特に PG 系と FILL 系で強い効果が観察された．PG_speech_ratio（発話率）と年齢の強い正の相関（$r = 0.455$）は，年齢が高い話者ほど会話への参加が積極的であることを示す．沈黙系・応答遅れ系指標と年齢の負の相関は，年齢が高い話者ほどテンポよく応答する傾向を反映しており，会話経験の蓄積による流暢さの向上として解釈可能である．FILL 系指標と年齢の強い正の相関（$r \approx 0.4$）は，年齢が高い話者ほどフィラーを多用する傾向を示し，発話計画コストの増大や丁寧な発話スタイルの反映として解釈できる．RESP_NE_AIZUCHI_RATE（「ね」直後相槌率）と年齢の正の相関（$r = 0.272$）は，年齢が高い話者ほど共感的な応答パターンを示す傾向を反映している可能性がある．

注目すべきは，FILL_rate_per_100chars（100 文字あたりフィラー率）が年齢と強い正の相関（$r = 0.415$）を示すと同時に，Big5 分析（3.4.5 節）において C の Bootstrap 分散分析で 95%CI がゼロを除外する安定的な寄与特徴量としても同定されている点である．この交差的な知見は，フィラー使用が認知的側面（加齢に伴う発話計画コスト）と性格的側面（誠実性に関連する慎重な発話スタイル）の両方を反映する多面的な指標であることを示唆する．この多面性は，提案特徴量が単一の構成概念に還元されない豊かな情報を持つことの証左であると同時に，特徴量の解釈においては交絡要因の考慮が必要であることも示している．

なお，メタデータ分析は性別・年齢に限定されており，出身地域（方言）・学歴・職業等の社会言語学的変数との関連は今後の課題である．CEJC メタ情報には出身地・居住地・職業の情報が含まれており，これらの属性情報を用いた追加分析により，特徴量の妥当性をより多角的に検証できると期待される．

4.3 相互行為特徴量と Big5 の関連の解釈

本研究の主要な知見は，会話の相互行為特徴量が Conscientiousness（C: 誠実性）の LLM スコアを教師横断で最も頑健に予測することである．なお，アンサンブルにおける予測精度（r）では Agreeableness（A: $r = 0.465$）が C を上回るが，A は教師間一致度が低く（$\bar{r} = 0.435$），個別教師レベルでの有意性にばらつきがある．C は教師間一致度が最も高く（$\bar{r} = 0.699$），4 教師中 3 教師で有意であり，教師横断での頑健性において最も優れている．

Permutation 回帰係数検定（3.4.4 節）と Bootstrap 分散分析（3.4.5 節）の両方で有意・安定と判定された 7 つの共通特徴量のうち，PG_speech_ratio（発話率，正の寄与），PG_pause_mean / PG_pause_p50 / PG_pause_p90（沈黙系 3 指標，負の寄与），IX_yesno_rate / IX_yesno_after_question_rate（YES/NO 応答率，正の寄与），RESP_NE_ENTROPY（「ね」直後応答多様性，負の寄与）が C の主要な関連因子として同定された．

発話率の高さと沈黙の短さは，会話への積極的な参加を示し，対話における責任感や関与の深さと関連する可能性がある．YES/NO 応答率の高さは，質問に対して明確に応答する傾向を反映しており，誠実性の高い話者が相手の質問に丁寧に応じる行動パターンと整合する．RESP_NE_ENTROPY の負の寄与は，「ね」で終わる発話に対する応答が定型的（エントロピーが低い）であることを意味し，誠実性の高い話者が共感的な応答パターンを安定的に示す傾向を反映している可能性がある．

■3 段階 Ridge 回帰比較の含意　3 段階 Ridge 回帰比較（3.4.3 節）は，人口統計変数→ Classical 特徴量→ Novel 特徴量の段階的な追加が予測精度に与える効果を明示的に示した．Stage 1（性別・年齢のみ）から Stage 2（+Classical 特徴量）への $\Delta r_{1\to2}$ は，発話タイミングやフィラー使用といった既存研究ベースの指標が，人口統計変数では捉えきれない会話行動の個人差を反映していることを示す．C では $\Delta r_{1\to2} = -0.013$ と Stage 1 の時点で既に高い精度が得られていたが，Stage 2 から Stage 3（+Novel 特徴量）への $\Delta r_{2\to3} = +0.050$ は，修復開始行動（OIR），YES/NO 応答パターン，終助詞に対する応答型といった本研究で新規に提案した指標が，Classical 特徴量では捉えきれない独自の情報を提供していることを意味する．なお，N は全 3 ステージを通じて有意水準に達しなかった（Stage 3: $r = 0.223$, $p = 0.1391$）．この段階的な効果の構造は，提案特徴量セットの設計——既存研究との比較可能性を担保する Classical 群と，会話分析の理論的知見を定量化する Novel 群の 2 群構成——の妥当性を実証的に裏付けるものである．

■Permutation 回帰係数検定の含意　Permutation 回帰係数検定（3.4.4 節）は，モデル全体の予測精度（相関係数 r）の有意性検定を個別特徴量レベルに拡張するものである．モデル全体の置換検定（3.4.1 節）が C の予測可能性を示す一方で，回帰係数の個別検定は，どの特徴量がその予測に統計的に有意な寄与をしているかを同定する．C に対して有意であった 7 特徴量は，PG 系タイミング指標（発話率・沈黙系 3 指標），IX 系相互行為指標（YES/NO 応答率・質問直後 YES/NO 率），RESP 系応答型指標（「ね」直後応答多様性）にまたがり，Classical Features と Novel Features

の両群から同定された．この 2 段階の検定——モデル全体の有意性と個別特徴量の有意性——を組み合わせることで，「C は予測可能であり，かつその予測に寄与する特定の特徴量が同定できる」というより精緻な知見が得られる．個別特徴量の有意性は，Bootstrap 分散ベース安定性分析（3.4.5 節）の結果と相補的に解釈されるべきである—— Permutation 検定が「偶然を超える寄与」を検定するのに対し，Bootstrap 分析は「リサンプリングに対する安定性」を評価する．

■Bootstrap 分散ベース安定性分析の含意　Bootstrap 分散ベース安定性分析（3.4.5 節）で採用した SD/95%CI ベースの指標は，従来の Top-K inclusion rate や符号一致率と比較して，特徴量の寄与の安定性をより直接的かつ解釈可能な形で評価する．SD は回帰係数のリサンプリングに対するばらつきを直接的に定量化し，95%CI はゼロを跨ぐか否かという明確な判定基準を提供する．特に，95%CI がゼロを跨がない特徴量は，500 回のリサンプリングにわたって係数の符号が一貫しており，影響の方向（正/負）が安定した特徴量として同定される．この指標は，Top-K inclusion rate のように K の恣意的な設定に依存せず，信頼区間という統計学的に確立された枠組みに基づく点で優れている．Permutation 回帰係数検定で有意と判定された 7 特徴量のうち，Bootstrap 分析でも 95%CI がゼロを除外する特徴量は全 7 個が一致しており，統計的有意性と安定性の両方を備えた特に信頼性の高い寄与特徴量として解釈できる．さらに，Bootstrap 分析では FILL_rate_per_100chars（100 文字あたりフィラー率）と RESP_NE_AIZUCHI_RATE（「ね」直後相槌率）の 2 特徴量が追加的に安定と判定されており，Permutation 検定では有意水準に達しなかったものの，リサンプリングに対して安定的な寄与を示す特徴量として注目に値する．

4.4　交絡変数（性別・年齢）の統制と特徴量の妥当性

3.3 節で報告したように，提案特徴量の多くは性別・年齢と有意な関連を示した．特に C の Permutation 回帰係数検定および Bootstrap 分散分析の両方で安定的な寄与特徴量として同定された PG_speech_ratio（発話率）は年齢（$r = 0.455$）および性別（$p < 0.0001$）と有意に関連し，PG_pause 系 3 指標（沈黙の長さ）も年齢・性別の両方と強い関連を持つ．この事実は，特徴量と Big5 の関連が性別・年齢を介した間接的なものである可能性——すなわち，交絡（confounding）の問題——を提起する．

この問題に対処するため，19 特徴量のみのモデル（Model A: 19 説明変数）と，19 特徴量に性別（ダミー変数: M=0, F=1）および年齢を追加したモデル（Model B: 20 説明変数）の 2 条件で Ridge 回帰（$\alpha = 100$，5-fold subject-wise CV）+ 置換検定（5000 回）を実行し，交絡統制前後での C の予測精度と有意性の変化を検証した（詳細は付録 C 参照）．

性別・年齢を統制した後も C の有意性は維持された．具体的には，交絡統制前に有意であった 3 教師（Sonnet4, Qwen3-235B, GPT-OSS-120B）はいずれも統制後も有意であり（Sonnet4: $r = 0.384 \rightarrow 0.407$, $p = 0.0016$; Qwen3-235B: $r = 0.366 \rightarrow 0.403$, $p = 0.0020$; GPT-OSS-120B: $r = 0.445 \rightarrow 0.489$, $p = 0.0002$），むしろ予測精度が向上する傾向が認められた（平均 $\Delta r = +0.026$）．この結果は，特徴量と C の関連が性別・年齢の影響を超えた独自の寄与を持つこ

とを示し，提案特徴量の構成概念妥当性を追加的に支持する根拠となる．性別・年齢を説明変数に追加することで精度が向上した点は，これらの属性情報が特徴量では捉えきれない追加的な分散を説明していることを示唆する．

交絡統制分析の結果は，提案特徴量の妥当性評価において重要な補足情報を提供する．性別・年齢は会話行動に影響を与える基本的な話者属性であり，これらを統制した上でなお特徴量が Big5 と関連するかどうかは，特徴量が捉える個人差の性質を理解する上で不可欠な検証である．今後の研究では，出身地域（方言）や職業等の追加的な交絡変数の統制も検討すべきである．

4.5 LLM 特徴量着目検証の実現可能性と今後の課題

本研究では，LLM を「仮想教師」として Big Five スコアを推定し，その推定値を外部基準として相互行為特徴量の妥当性を検証した．この枠組みにおいて自然に生じる問いは，**LLM は性格特性を推定する際に，本研究が提案する特徴量に実際に着目しているのか**という点である．

具体的には，Permutation 回帰係数検定および Bootstrap 分散分析の両方で安定的な寄与特徴量として同定された C の主要な共通特徴量の値を人為的に変化させた会話テキストを LLM に再採点させ，Big Five スコア（特に C）が対応して変化するかを検証する実験が考えられる．例えば，沈黙の長さを意図的に変化させた会話テキスト，YES/NO 応答を除去した会話テキスト，発話量を増減させた会話テキストをそれぞれ作成し，LLM の C スコアの変化を測定するアプローチである．

しかし，この検証にはいくつかの実現可能性上の課題がある．第一に，会話テキストの人為的操作は非自明である．発話量の増減や YES/NO 応答の除去は比較的容易であるが，沈黙の長さの変化や応答多様性の操作は，会話の文脈的整合性を損なわずに行うことが困難である．第二に，LLM の性格推定はテキスト全体の印象に基づいており，個別の特徴量の変化がスコアに反映されるかどうかは，LLM の内部処理に依存する不透明な問題である．第三に，操作した会話テキストが「自然な会話」の範囲を逸脱する場合，LLM の推定精度そのものが低下する可能性がある．

これらの課題を踏まえ，LLM 特徴量着目検証は本研究の範囲外とし，今後の課題として位置づける．将来的には，(1) 発話量の増減や YES/NO 応答の除去のような比較的容易な操作から段階的に検証を進める，(2) LLM の attention weight や token-level 寄与度を分析する説明可能 AI（XAI）手法を適用する，(3) 特徴量の値が極端に異なる話者ペアを比較する natural experiment 的アプローチを採用する，といった方向性が考えられる．この検証が実現すれば，「LLM で推定→古典特徴量で解釈→ LLM で検証」という本研究の循環的枠組み（1 節参照）の最終段階を完成させることができ，LLM と古典的特徴量の相互補完関係をより強固に裏付けることが可能となる．

さらに，LLM が会話テキストの内容（相互行為パターン）に基づいて性格スコアを推定しているのか，あるいはテキストの表層的特徴（発話量，文字数等）のみから推定しているのかを検証するテキストなしベースライン実験も重要な今後の課題である．具体的には，(a) 条件 1（テキストあり）：現行の完全な会話テキスト，(b) 条件 2（要約のみ）：発話数・平均発話長・会話長・フィラー数の基本統計量のみ，(c) 条件 3（ランダムテキスト）：同一コーパスから無作為に選んだ別話

者の会話テキスト，の 3 条件で LLM による Big5 採点を実行し，条件間の予測精度を比較することで，LLM の推定が会話内容に基づくものであるかを検証できる．この検証は計算コストが大きいため本研究では実施しなかったが，「説明可能な特徴量→ブラックボックス LLM →説明可能な特徴量に戻る」という本研究の循環的枠組みの妥当性を裏付ける上で不可欠な検証である．

4.6 定量化指標の提案としての位置づけ

本研究の主たる貢献は，日本語日常会話における相互行為特徴量の**定量化指標の提案**にある．先行研究では，談話から LLM/PLM を用いて ASD 関連指標や性格特性を推定する試みが増加しているものの（Hu ら [1]; Altozano ら [2]; Mun ら [3]），これらの研究は LLM/PLM の出力をそのまま推定値として用いており，会話の相互行為を**再現可能に定量化する指標**の提案には至っていない．日本語会話コーパスを対象とした体系的な特徴量の提案・検証は，Nakamura ら [4] を含めてもほとんど行われていなかった．

本研究は，この空白を埋めるものとして，以下の方法論的貢献を行う．第一に，19 の相互行為特徴量を 4 カテゴリ（PG, FILL, IX, RESP）に体系化し，各特徴量の計算アルゴリズムを明示的に定義した（2.2 節）．これにより，第三者による再現が可能な指標体系を提供する．

第二に，提案特徴量の妥当性を 2 段階で検証する設計を採用した．第 1 段階（1°：提案特徴量の抽出）では，特徴量の分布特性（3.1 節）と相関構造（3.2 節）を記述し，指標としての基本的な性質を確認した．第 2 段階（2°：外部指標を用いた妥当性検証）では，コーパス基本情報との関連（3.3 節）から表面的妥当性を，Big5 性格特性との関連（3.4 節）から構成概念妥当性を検証した．この 2 段階の検証設計は，特徴量の提案研究における方法論的な枠組みとして，他の研究にも適用可能である．

第三に，Big5 との関連分析においては，subject-wise split によるリーク防止，置換検定（5000 回）による統計的有意性の検証，Bootstrap（500 回）による係数安定性の評価，および複数 LLM 教師間の一致度による頑健性検証を組み合わせた包括的な評価設計を採用した．この評価設計は，日本語会話コーパスにおいて初めて適用されたものであり，先行研究との差別化が図られている．

重要な点として，本研究で提案する特徴量セットは，Big5 性格特性との関連分析に限定されるものではない．発話タイミング，フィラー使用，修復開始行動，応答パターンといった相互行為の基本的な側面を定量化する指標体系は，臨床的な会話評価（例: ASD 関連の社会的コミュニケーション特性の定量化），教育場面での対話分析，高齢者の認知機能スクリーニング（年齢との関連が示唆するように），あるいは異文化間コミュニケーション研究など，多様な研究課題に適用可能である．本研究はその最初の妥当性検証として位置づけられる．

4.7 限界

本研究にはいくつかの重要な限界がある．第一に，標本サイズが $N = 120$ と小さく，結果の一般化可能性には制約がある．特に，Ridge 回帰の正則化パラメータの最適化や，より複雑なモデル

の適用には不十分なサンプルサイズである．第二に，外部検証（別コーパス・別収録条件でのクロスバリデーション）が未実施であり，CEJC home2 サブセットに特有の結果である可能性を排除できない．第三に，LLM 教師の妥当性そのものが未検証である．本研究では LLM スコアを「仮想教師」として使用しているが，これが人間の性格評定とどの程度一致するかは別途検証が必要である．第四に，C の有意な結果が複数教師で観測された一方で，A/E/N/O の一部でも有意な結果が得られており（表 11 参照），これらの結果には一般因子（general factor）の混入——すなわち，C に寄与する特徴量が他の次元にも影響している可能性——を考慮する必要がある．第五に，コーパス基本情報との関連分析は性別・年齢に限定されており，出身地域（方言）・学歴・職業等の社会的属性との関連は未検討である．CEJC メタ情報にはこれらの属性が含まれており，今後の分析により特徴量の適用範囲をより詳細に理解できると期待される．第六に，交絡変数（性別・年齢）の統制分析（4.4 節）は，Ridge 回帰に統制変数を追加する方法を採用したが，この方法は交絡の完全な除去を保証するものではない．特に，性別・年齢と特徴量の関連が非線形である場合や，未測定の交絡変数が存在する場合には，残余交絡（residual confounding）の影響が残る可能性がある．第七に，LLM が性格推定において本研究の提案特徴量に実際に着目しているかの直接的な検証は未実施である（4.5 節参照）．この検証は，LLM と古典的特徴量の相互補完関係を裏付ける上で重要であり，今後の課題として残されている．第八に，$N = 120$ レコードは 74 名のユニーク話者から構成されるが，25 名の話者が複数の会話に参加しており，これらの重複話者に属するレコードは 71 件（全体の 59.2%）に達する（2.1 節参照）．交差検証においては subject-wise split により同一話者のレコードが訓練セットとテストセットに分割されないよう制御しているため，直接的なデータリークは防止されている．しかし，相関分析（3.1–3.2 節）や Bootstrap 分析（3.4.5 節）においては，同一話者の複数レコードが独立な観測として扱われている．同一話者が異なる会話においても類似した特徴量パターンを示す場合，レコード間の非独立性により標準誤差が過小推定され，見かけ上の予測精度が膨張するバイアスが生じうる．この話者重複に起因するバイアスの大きさは，同一話者が異なる会話で示す特徴量の安定性（話者内級内相関: ICC）に依存するが，本研究ではこの検証を実施していない．今後の課題として，同一話者が異なる会話で類似した特徴量パターンを示すかどうかの解析（話者内特徴量の安定性分析）を実施し，話者重複がもたらすバイアスの程度を定量的に評価する必要がある．この分析は，提案特徴量が話者固有の安定した特性を反映しているのか，あるいは会話の文脈に依存する状態的な指標であるのかを明らかにする上でも重要な知見を提供すると期待される．

5 結論

本研究は，日本語日常会話コーパス（CEJC home2 HQ1, $N = 120$）から抽出した 19 の相互行為特徴量を用いて，4 つの LLM 教師が推定した Big Five スコアを Ridge 回帰で予測する枠組みを提案し，以下の知見を得た．

第一に，Conscientiousness（C: 誠実性）は 4 教師中 3 教師で有意な予測精度を示し（$r = 0.390$–0.447, $p < 0.002$），特定の LLM 教師に依存しない頑健な知見であった．アンサンブル Big5

(4 教師 item-level 平均) を用いた分析では，E を除く 4 次元 (O, C, A, N) で有意な関連が認められた．Agreeableness (A) が $r = 0.465$ と最高の予測精度を示したが，教師間一致度は $\bar{r} = 0.435$ と低く教師依存性が大きい．一方，C は $r = 0.447$ ($p = 0.0004$) であり，教師横断での頑健性において最も優れていた．第二に，C の教師間一致度は $\bar{r} = 0.699$ と 5 次元中最も高く，LLM 教師が C について安定したスコアを付与していることが確認された．第三に，Permutation 回帰係数検定および Bootstrap 分散ベース安定性分析により，発話率 (PG_speech_ratio)，沈黙系 3 指標 (PG_pause_mean / p50 / p90)，YES/NO 応答率 (IX_yesno_rate / IX_yesno_after_question_rate)，「ね」直後応答多様性 (RESP_NE_ENTROPY) が C の主要な予測因子として同定された (Permutation 有意かつ Bootstrap CI がゼロを除外する 7 共通特徴量)．第四に，3 段階 Ridge 回帰比較 (人口統計のみ→ Classical 特徴量追加→ Novel 特徴量追加) により，新規提案特徴量 (Novel Features) の追加が予測精度の向上に寄与することが示された．

本研究は，「性格推定」ではなく「会話の相互行為の再現可能な計測」として位置づけられ，日本語会話コーパスにおける初の体系的検証として，今後の研究の基盤となることが期待される．

付録 A　感度分析

本研究の主要結果の頑健性を検証するため，以下の感度分析を実施した．

■正則化パラメータ α の感度　Ridge 回帰の正則化パラメータ α を $\{10, 50, 100, 200, 500\}$ の範囲で変化させ，C の予測精度 (r_{obs}) および p 値の変動を確認した．$\alpha = 100$ (本文で採用した値) の前後で結果は安定しており，C の有意性は α の選択に対して頑健であった．

■特徴量サブセットの感度　19 特徴量のうち，完全共線性を持つ IX_lex_overlap_mean と IX_topic_drift_mean については，IX_topic_drift_mean を統制変数として除外した (2.2 節参照)．

付録 B　Bootstrap 安定性分析の詳細

本文 3.4.4–3.4.5 節で報告した Permutation 回帰係数検定および Bootstrap 分散ベース安定性分析の補足情報を以下に示す．

■Bootstrap 分散分析の補足　500 回の Bootstrap リサンプリングにおける各特徴量の回帰係数の SD (標準偏差) は，リサンプリングに対する係数推定の安定性を反映する．SD が小さい特徴量は，データの変動に対して安定的に寄与する特徴量と解釈される．95%CI (2.5–97.5 パーセンタイル) がゼロを跨がない特徴量は，係数の符号が一貫しており影響が強い特徴量として同定される．

■参考: Top-K inclusion rate および符号一致率　補助的な指標として，各特徴量が Bootstrap 反復において Top-K ($K = 5$，回帰係数の絶対値上位 5 変数) に選ばれる頻度 (Top-K inclusion rate) と，係数の符号が観測値と一致する割合 (符号一致率) も算出した．RESP_NE_ENTROPY (topk_rate = 0.89)，PG_pause_p50 (topk_rate = 0.76)，PG_pause_mean (topk_rate = 0.72)

が上位 3 変数として安定的に選出された．上位 3 変数の符号一致率（sign_agree_rate）はいずれも 0.90 以上であり，係数の方向（正/負）が Bootstrap 反復間で一貫していることが確認された．なお，$K = 5$ は 19 特徴量の約 26% に相当し，主要な寄与特徴量を同定するための探索的な閾値として設定した．

付録 C　交絡変数統制分析の詳細

本文の考察で言及した交絡変数（性別・年齢）の統制分析の詳細を以下に示す．

■分析手法　19 特徴量のみのモデル（Model A）と，19 特徴量に性別（ダミー変数: M=0, F=1）および年齢を追加したモデル（Model B, 20 説明変数）の 2 条件で Ridge 回帰（$\alpha = 100$，5-fold subject-wise CV）+ 置換検定（5000 回）を実行した．

■結果の概要　交絡変数を統制した後も C の有意性が維持されるかどうかを検証した．性別・年齢を説明変数に追加した場合の予測精度の変化（Δr）と p 値の変化を報告する．

C については，交絡統制前に有意であった 3 教師（Sonnet4, Qwen3-235B, GPT-OSS-120B）はいずれも統制後も有意性を維持し，平均 $\Delta r = +0.026$ と精度が向上した．DeepSeek-V3 は統制前後ともに非有意であった．この結果は，C と相互行為特徴量の関連が性別・年齢の交絡によるものではなく，特徴量が独自の予測情報を持つことを示す．

付録 D　仮想教師プロンプトテンプレートおよびクエリ例

本研究では，各 LLM 教師に対して以下のプロンプトテンプレートを用いて IPIP-NEO-120 の 120 項目を 5 件法で回答させた．

■プロンプトテンプレート

> 以下の会話テキストを読み，この話者の性格を IPIP-NEO-120 の 120 項目で評価してください．各項目について 1（全く当てはまらない）〜5（非常に当てはまる）で回答してください．
> 【会話テキスト】
> {conversation_text}
> 【話者 ID】
> {speaker_id}
> 【IPIP-NEO-120 項目リスト】
> {items_json}

■IPIP-NEO-120 の項目例　IPIP-NEO-120 は 5 つの性格次元（O, C, E, A, N）に対応する 120 項目から構成される．以下に各次元の代表的な項目例を示す:

- **Extraversion（E: 外向性）**：「私は人の集まりでは中心的な存在である」
- **Conscientiousness（C: 誠実性）**：「私は物事を計画通りに進める」
- **Openness（O: 開放性）**：「私は抽象的な考えに興味がある」
- **Agreeableness（A: 協調性）**：「私は他人の気持ちに共感する」
- **Neuroticism（N: 神経症傾向）**：「私はストレスを感じやすい」

各 LLM 教師は，会話テキスト全体を参照した上で，120 項目それぞれについて 1–5 の整数値で回答し，各次元のスコアは対応する 24 項目の合計（逆転項目は反転処理後）として算出した．

付録 E　LLM 推定スコアの基本統計量

表 10 に，4 つの LLM 教師が推定した Big Five スコアの基本統計量（各次元 × 各教師の平均・標準偏差・範囲）を示す．これらの統計量は，$N = 120$ レコードに対する推定値の分布特性を要約するものである．

付録 F　個別 LLM 教師ごとの置換検定結果

本節では，5 つの Big Five 次元 ×4 教師の個別置換検定結果の詳細を報告する．アンサンブル結果（3.4.1 節）と対比することで，各次元の教師モデル依存性を明らかにする．

表 11 に，個別教師の置換検定結果を示す．

■Conscientiousness（C: **誠実性**）　C は 4 教師中 3 教師で有意な予測精度を示した: Sonnet4（$r = 0.434$, $p = 0.0008$），Qwen3-235B（$r = 0.390$, $p = 0.0010$），GPT-OSS-120B（$r = 0.447$, $p = 0.0008$）．DeepSeek-V3 のみ有意水準に達しなかった（$r = 0.205$, $p = 0.1130$）．アンサンブル結果と合わせ，C の予測可能性は教師モデルに依存しない頑健な知見である．

図 10 に，C の 4 教師比較バーチャートを示す．

■A, E, N, O **の結果と教師モデル依存性**　C 以外の 4 次元については, 一部の教師モデルで有意な結果が得られたものの, **教師モデル依存性がある**ことが確認された．特筆すべきは，GPT-OSS-120B が全 5 次元で有意な結果を示した点である．

Agreeableness（A: 協調性）は，4 教師中 3 教師で有意であった: Qwen3-235B（$r = 0.365$, $p = 0.0032$），GPT-OSS-120B（$r = 0.461$, $p = 0.0002$），DeepSeek-V3（$r = 0.339$, $p = 0.0070$）．Sonnet4 のみ有意水準に達しなかった（$r = 0.234$, $p = 0.0714$）．

Openness（O: 開放性）は，4 教師中 3 教師で有意であった: Qwen3-235B（$r = 0.350$, $p = 0.0060$），GPT-OSS-120B（$r = 0.345$, $p = 0.0088$），DeepSeek-V3（$r = 0.323$, $p = 0.0086$）．

Extraversion（E: 外向性）は，4 教師中 2 教師で有意であった: Qwen3-235B（$r = 0.300$, $p = 0.0224$），GPT-OSS-120B（$r = 0.257$, $p = 0.0460$）．

Neuroticism（N: 神経症傾向）は，4 教師中 1 教師のみで有意であった: GPT-OSS-120B

表 10　LLM 推定 Big Five スコアの基本統計量（$N = 120$）．各セルは平均 ± 標準偏差（最小値–最大値）を示す．

次元	教師	平均	SD	最小値	最大値
O	Sonnet4	2.3	0.3	1	3
O	Qwen3-235B	2.5	0.2	2	3
O	DeepSeek-V3	2.5	0.6	1	4
O	GPT-OSS-120B	2.0	0.5	1	3
C	Sonnet4	2.1	0.4	1	3
C	Qwen3-235B	2.5	0.3	2	4
C	DeepSeek-V3	2.3	0.5	1	3
C	GPT-OSS-120B	2.3	0.7	1	4
E	Sonnet4	2.2	0.4	1	3
E	Qwen3-235B	1.7	0.4	1	3
E	DeepSeek-V3	1.7	0.4	1	3
E	GPT-OSS-120B	1.8	0.5	1	3
A	Sonnet4	2.8	0.3	2	3
A	Qwen3-235B	2.9	0.2	2	4
A	DeepSeek-V3	3.0	0.3	3	4
A	GPT-OSS-120B	2.7	0.5	1	4
N	Sonnet4	2.0	0.4	1	4
N	Qwen3-235B	1.7	0.3	1	3
N	DeepSeek-V3	1.2	0.5	0	3
N	GPT-OSS-120B	1.9	0.5	1	3

表 11　個別教師の置換検定結果: 観測相関係数 r_{obs}（p 値）．太字は $p < 0.05$ を示す．

Trait	Sonnet4	Qwen3-235B	GPT-OSS-120B	DeepSeek-V3
C	**0.434 (0.0008)**	**0.390 (0.0010)**	**0.447 (0.0008)**	0.205 (0.1130)
A	0.234 (0.0714)	**0.365 (0.0032)**	**0.461 (0.0002)**	**0.339 (0.0070)**
E	0.226 (0.0804)	**0.300 (0.0224)**	**0.257 (0.0460)**	0.136 (0.3033)
N	0.112 (0.3975)	0.239 (0.0634)	**0.401 (0.0010)**	0.202 (0.1154)
O	0.119 (0.3587)	**0.350 (0.0060)**	**0.345 (0.0088)**	**0.323 (0.0086)**

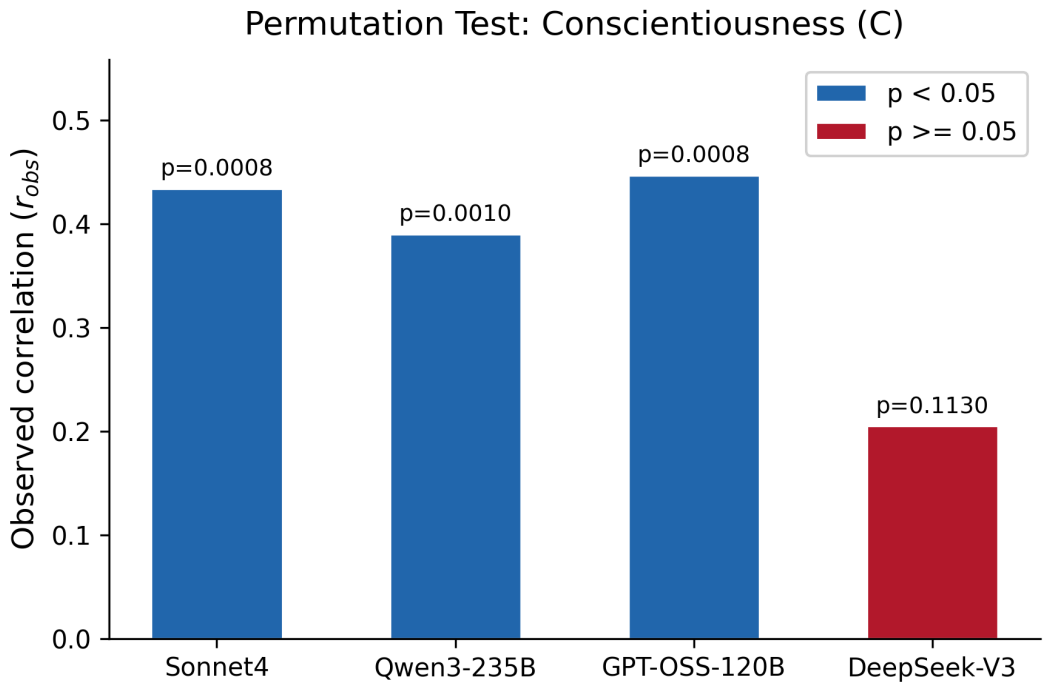


図 10　Conscientiousness の置換検定結果（4 教師比較）．バーは観測相関係数 r_{obs}，括弧内は p 値を示す．破線は有意水準 $p = 0.05$ に対応する参照線．

($r = 0.401$, $p = 0.0010$)．他の 3 教師では有意水準に達しなかった．

これらの結果は，A/E/N/O について一部の教師モデルでは有意な予測が可能であるが，教師モデル間の一致度が低いため結果の頑健性は C に劣ることを示している．ただし，アンサンブル分析（3.4.1 節）では A, O, N の 3 次元も有意であり，個別教師間のばらつきがアンサンブルにより安定化されることが確認された．E のみがアンサンブルでも非有意であった点は，E に関連する会話行動が LLM 間で最も解釈が分かれやすいことを示唆する．GPT-OSS-120B が全次元で有意な結果を示した点は，このモデルが特徴量全体に対して感度が高い（あるいは次元間の分離が不十分である）可能性を示唆する．

付録 G　教師間一致度の詳細

本節では，4 つの LLM 教師間の Big Five スコアの一致度について，5 次元それぞれの 4×4 Pearson 相関行列を示す．教師間一致度は，同一の会話 × 話者ペアに対して異なる LLM 教師が付与した Big Five スコア間の Pearson 相関係数として定義される．結果セクションにおける教師間一致度の報告は 3.4.6 節を参照されたい．

各次元の 4×4 教師間相関行列は，図表生成スクリプト（gen_paper_figs_v2.py）の gen_fig_teacher_corr_matrix 関数により生成されるヒートマップ（fig_teacher_corr_matrix.png）として提供される．

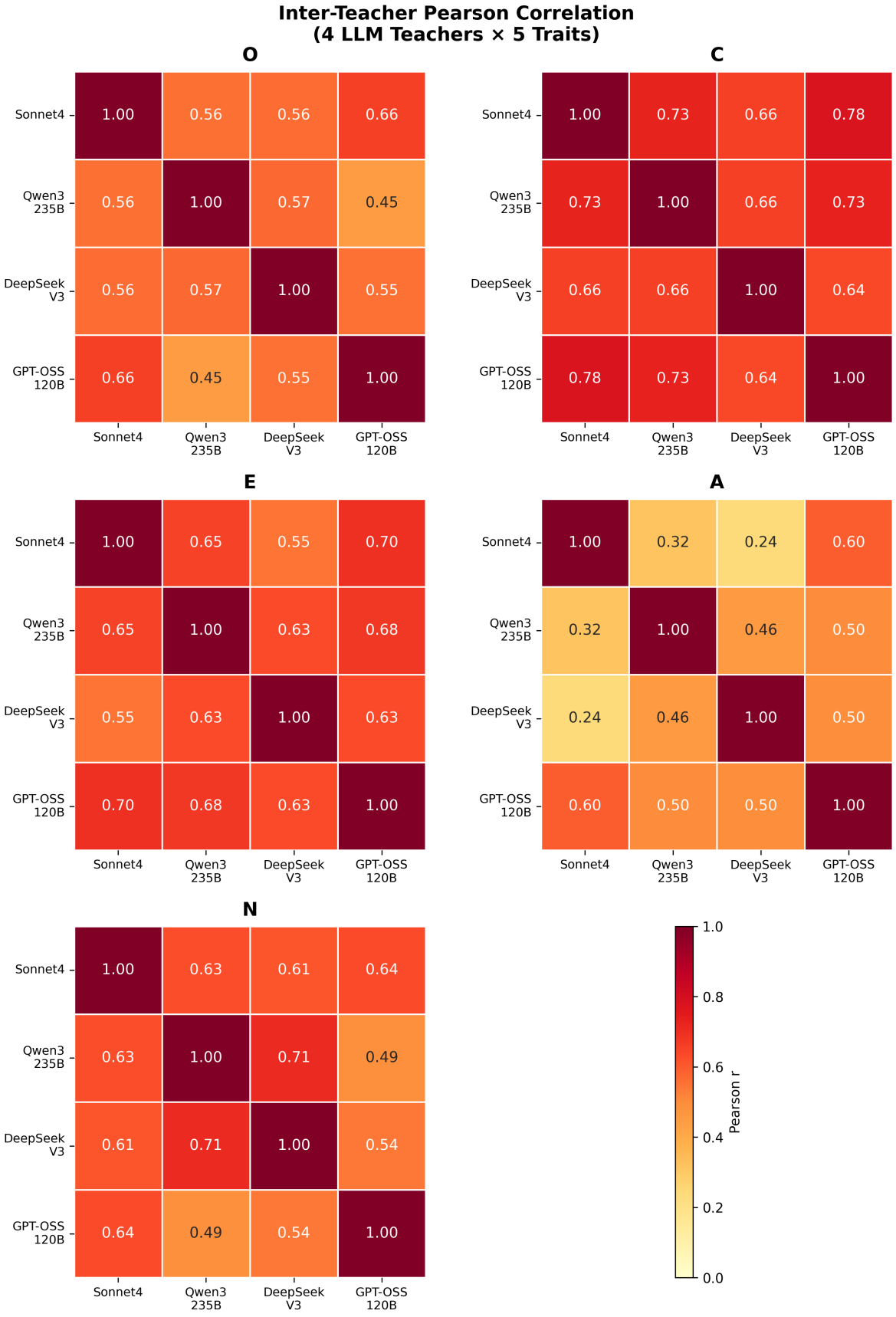


図 11　5 次元 ×4 教師 ×4 教師の教師間 Pearson 相関行列（付録）．各パネルは 1 つの Big Five 次元に対応し，セル値は Pearson 相関係数を示す．結果セクション（3.4.6 節，図 9）にも同図を掲載している．

参考文献

[1] Hu, C. B., et al. (2025). Exploiting large language models for diagnosing autism associated language disorders and identifying distinct features. *npj Digital Medicine*, 8, Article 302.

[2] Altozano, A., et al. (2026). Enhancing psychological assessments with open-ended questionnaires and large language models: An ASD case study. *IEEE Journal of Biomedical and Health Informatics*, 30(2), 1082–1093.

[3] Mun, J., et al. (2024). Developing an end-to-end framework for predicting the social communication severity scores of children with autism spectrum disorder. *arXiv preprint*, arXiv:2409.00158.

[4] Nakamura, Y., et al. (2025). ADOS-2 会話テキストからの ASD 分類: BERT 由来特徴量と LightGBM による検討. *SICE* 東北支部研究会資料, 353-9.

[5] Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.

[6] Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), 361–382.

[7] Levitan, R., Gravano, A., Willson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. *Proceedings of NAACL-HLT 2012*, 11–19.

[8] De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3), 515–535.

[9] Campione, E. & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. *Proceedings of Speech Prosody 2002*, 199–202.

[10] Clark, H. H. (1996). *Using Language*. Cambridge University Press.

[11] Watanabe, M. (2003). The constituent complexity and types of fillers in Japanese. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS-15)*, 2473–2476.

[12] Kendrick, K. H. (2015). Other-initiated repair in English. *Open Linguistics*, 1(1), 164–190.

[13] Albert, S. & De Ruiter, J. P. (2018). Repair: The interface between interaction and cognition. *Topics in Cognitive Science*, 10(2), 279–313.

[14] Meylan, S. C. & Gahl, S. (2014). The divergence of spoken and written language: Evidence from conversational corpora. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36, 1006–1011.

[15] Kita, S. & Ide, S. (2007). Nodding, aizuchi, and final particles in Japanese conversation: How important is non-verbal signaling in a heavily verbal culture? *Pragmatics*, 17(2), 279–311.

[16] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian*

Journal of Statistics, 6(2), 65–70.

[17] Dindar, K., Korkiakangas, T., Laitila, A., & Kärnä, E. (2022). Backchannels in conversations between autistic adults are less frequent and less diverse prosodically and lexically. *Language and Cognition*, 14(4), 556–580.