



## L2 英語学習者による説得型ロールプレイ発話の評価 ：評価指標間の関係性の解明と発話スタイル別サン プルの検討

石川， 慎一郎

---

(Citation)

統計数理研究所共同研究リポート, 465:78-91

(Issue Date)

2023-03-10

(Resource Type)

departmental bulletin paper

(Version)

Version of Record

(JaLCD0I)

<https://doi.org/10.24546/0100479388>

(URL)

<https://hdl.handle.net/20.500.14094/0100479388>



## L2 英語学習者による説得型ロールプレイ発話の評価

### —評価指標間の関係性の解明と発話スタイル別サンプルの検討—

石川 慎一郎(神戸大学)

iskwshin@gmail.com

Assessment of L2 English Learner Utterances in Persuasion Roleplays  
—Consideration of the Relationship between Assessment Indices and  
Identification of Learner Samples Linked to Particular Speech Styles—  
ISHIKAWA, Shin'ichirio (Kobe University)

#### Abstract

Using the data from the ICNALE Global Rating Archives (Ishikawa, 2020), this study examined the ratings that 65 raters assigned to the utterances of 140 participants consisting of Asian L2 English learners as well as English native speakers during the persuasion roleplays. Analytical foci were put on (1) the relationship between 13 kinds of assessment indices (10 kinds of analytical assessment scores, a holistic assessment score, an analytical assessment score sum, and an overall rating score), (2) statistical loadings of the analytical assessment indices, and (3) identification of the speech samples representing different speech styles. The quantitative analyses showed that (1') 13 indices were divided into two, which are related to language-oriented speeches in which speakers pay attention to the linguistic quality of their speeches and content-oriented speeches in which they pay attention to logicity and purposefulness in their speeches and show more willingness to communicate and intention to involve hearers; (2') roughly the same loadings were given to different analytical assessment indices; and (3') four speech styles were identified, which were observed more clearly in the speeches of EFL learners at novice and lower intermediate levels.

#### キーワード

発話評価、パフォーマンス評価、評価観点、学習者コーパス、評価データセット

#### 1. はじめに

「できる(であろう)こと」と「実際にやっていること」の間には、往々にして壁が存在する。L2 習得に関して言えば、習熟度の高い学習者であれば、常に質の高い産出を行うものと予想されるわけ

だが、実態は必ずしもそうではない。習熟度テストの高得点者が、実際にほとんど話せなかったり、中身のある作文を書けなかったりする例は決して珍しいことではない。このことは、習得の問題を論じるにあたり、潜在的能力を示す習熟度と、顕在的能力を示す産出の質の両面を考えていく必要があることを示唆する。

L2 習得を量的に議論しようとする場合、様々な背景属性を持つ学習者の産出を大規模に集めた学習者コーパスの活用が有益であるが、習熟度と産出の質の両面のデータを有するコーパスはきわめて少ない。これは、筆者が構築しているアジア圏国際英語学習者コーパス ICNALE (International Corpus Network of Asian Learners of English) にもあてはまる問題であった。アジア圏学習者による 15,000 件以上の L2 産出サンプルを収録する ICNALE は、TOEIC や TOEFL といった習熟度テストのスコア、または、共通して実施した語彙テストのスコアに基づき、学習者ごとの習熟度を 4 段階に区分していたが、産出サンプル自体の質を表す情報はなかった。

そこで、2020 年度に始まった ICNALE Global Rating Archives (ICNALE GRA) プロジェクトでは、収集済みのサンプルの中から、アジア圏の英語学習者および英語母語話者による 140 本の発話(説得型ロールプレイの冒頭部分)と同数の作文を選び、それぞれ 60 名以上の多様な背景を有する評価者から評価データを取得した。

ICNALE GRA は、既存の評価データセット以上に、多数の、かつ、幅広い背景を持つ評価者から、統制的な方法で評価データが収集されている。そのため、幅広い評価者の集合知に基づいて、各レベルの代表的なサンプルを抽出することも可能になる。Ishikawa (2023)では、60 名の評価者による全体的評価(holistic assessment)スコアと、10 観点別の分析的評価(analytic assessment)スコアの合計値を平均した統合評価指標(overall rating score: ORS)に基づき、各サンプルを A(ORS が 70%以上)、B(60%以上)、C(50%以上)、D(40%以上)、E(30%以上)、F(30%未満)に分類した。また、各レベルに分類されたサンプルの中で、変動係数(coefficient of variation)が最小になる(つまり、評価値のばらつきが極小となる)4 サンプルを選び、各レベルを代表するベンチマークサンプルとして特定した。L2 産出の評価は、ハイステイクステストであっても、1 名ないし若干名の判断で決定される場合が多いことをふまえると、60 名以上の評価者の集合知を根拠とするベンチマークサンプルは、レベル別の学習者言語の特性研究や、L2 指導の指針として、幅広い活用が期待されるものである。

ただし、Ishikawa (2023)では、ORS の算出に先立って 10 観点別の分析的評価スコアを単純合計したわけだが、この処理には検討の余地もあろう。というのも、10 観点を単純合算する場合、10 観点のすべてに均等のウェイトをかけたことになるが、仮に、観点の中に著しく情報量が多い、または乏しいものが存在していた場合、あるいは、事実上同じ内容を測定している複数の観点が存在していた場合、単純合算の妥当性は低減するからである。

そこで、本研究では、作文に比べ、評価が安定しにくい発話サンプルを対象を絞り、10 観点別の分析的評価スコアと、それらの合成で得られた 3 種のスコア、あわせて 13 種のスコアをそれぞれ独立した変数とみなした上で、これらの関係性を確認する。また、10 観点別の分析的評価スコアに関して、2 種類の多変量解析手法で変数圧縮を行い、代表変数における各観点のウェイトが、

Ishikawa (2023)で採用した均等ウェイトからどの程度乖離しているかを検証する。最後に、Ishikawa (2023)で提示したレベル別のベンチマークサンプルを拡張するものとして、新たに、発話スタイル別のベンチマークサンプルの特定を試みる。

なお、本研究に関連する論考として、Ishikawa (2020)では、ICNALE GRA プロジェクトの目的と概要を述べた。また、Ishikawa(2023)は、ICNALE プロジェクトの全体像と活用例をまとめた概説書で、3.6 節では ICNALE GRA の構築過程の詳細を紹介した。加えて、執筆時点で収集が完了していたデータに基づき、8.2 節では評価データの信頼性検証、8.3 節では言語指標に基づく産出品質の自動推定、8.4 節ではベンチマークサンプル抽出のテーマを論じた。Ishikawa (in press-a)では、中国・日本・タイの学習者の発話産出に対する評価データを用い、評価者の母語、評価経験、英語教授経験が評価スコアに及ぼす影響について検証した。Ishikawa (in press-b)では、学習者コーパス研究における中間言語対照分析手法の再考を目指す観点から、従来の研究が一般的に依拠していた母語話者の産出サンプルではなく、ICNALE GRA で抽出された A レベルの学習者サンプルを新しい基準とした場合に、学習者の特徴的語彙使用(過剰使用・過少使用)がどのように変化するかを論じた。

## 2. リサーチデザイン

### 2.1 目的と研究設問

本研究の目的は、ICNALE GRA で収集された評価データのうち、発話評価のデータを再分析することで、10 観点別の分析的評価スコアの観点間の関係性と、Ishikawa (2023)で行われたスコア合成手法の妥当性を検証し、発話スタイル別ベンチマークサンプルの抽出を行うことである。具体的には以下の 3 つの研究設問(RQ)の解明を目指す。

RQ1 13 種の評価指標はどのような相互関係にあるか？

RQ2 10 観点別の分析的評価スコアを多変量解析手法で合成した場合、評価観点別のウェイトはどのように変化するか？

RQ3 どのような発話スタイルがあり、それらを代表するサンプルはどのようなものか？

### 2.2 データ

本研究では、ICNALE GRA の評価データから、140 名の話者によるロールプレイ内発話に対する 65 名の評価者の評価データを使用する。なお、Ishikawa (2023)では 60 名分の評価者データを使用していたが、その後に増えた 5 名のデータを今回の分析に加えている。

#### 2.2.1 発話者

140 名の話者のうち、非母語話者は全員が大学生(大学院生を含む)で、属性は表 1 のとおりである。学習者の習熟度は、TOEIC や TOEFL などの習熟度テストのスコア、または全参加者対象に受験を義務付けた共通語彙サイズテストのスコアに基づき、Common European Framework of Reference for Languages (CEFR)が定める 4 つの段階(A2～B2+)に区分されている。

表 1

話者 140 名の地域・習熟度のプロフィール

| 圏域    | 地域(略号)      | A2 | B1_1 | B1_2 | B2+ | 小計  |
|-------|-------------|----|------|------|-----|-----|
| EFL 圏 | 中国(CHN)     | 2  | 6    | 6    | 6   | 20  |
|       | インドネシア(IDN) | 6  | 6    | 6    | 2   | 20  |
|       | 日本(JPN)     | 5  | 5    | 5    | 5   | 20  |
|       | 韓国(KOR)     |    | 3    | 7    | 10  | 20  |
|       | タイ(THA)     | 6  | 6    | 6    | 2   | 20  |
|       | 台湾(TWN)     | 5  | 5    | 5    | 5   | 20  |
|       | 香港(HKG)     |    |      |      | 4   | 4   |
| ESL 圏 | パキスタン(PAK)  |    |      | 3    | 1   | 4   |
|       | フィリピン(PHL)  |    |      |      | 4   | 4   |
|       | マレーシア(MYS)  |    |      | 4    |     | 4   |
| 母語話者  | 英語母語話者(ENS) |    |      |      |     | 4   |
|       | 小計          | 24 | 31   | 42   | 39  | 140 |

英語を外国語として学んでいる EFL(English as a foreign language)圏については 1 地域 20 名のサンプルを取った。できるだけ各習熟度のサンプルがバランスよく含まれるよう配慮したが、韓国では A2 レベルの参加者がいなかったため、また、インドネシアでは B2+ レベルの参加者が少なかったため、一部の地域においては若干の人数の偏りが出ている。英語を公用語または同等の位置づけに置く ESL(English as a second language)圏と母語話者については、参考データと位置づけ、それぞれ 4 名ずつのサンプルを取った。

## 2.2.2 発話タスク

評価対象としたのは、ICNALE Spoken Dialogues(Ishikawa, 2019)で収集されたインタビューデータ中のロールプレイ発話である。このコーパスには 2 種のロールプレイタスクが含まれているが、ICNALE GRA で評価したのはアルバイトをテーマとしたロールプレイである。インタビューに参加した学習者は、学生は学業に専念すべきだと考えている指導教員に、自身が行っているアルバイトの継続を認めてもらえるよう説得を行う、という指示を与えられた。

ロールプレイの中で、学習者は、主体的に議論を先導・展開し、教員役のインタビュワーを説得する必要がある。このため、本タスクでは、表層的な L2 の発話力だけでなく、語用論的能力や、対人関係のコントロール能力、さらには、説明や提案の構想力など、実際の口頭コミュニケーション場面で必要となる多面的な能力が要求されることとなる。

ロールプレイは数分間に及ぶが、冒頭の 90 秒を一律で切り出して評価サンプルとした。すべてのサンプルは、話者属性を秘匿した上で、音声ファイルの形で評価者に提供された。各サンプルにはインタビュワー側の発話も入っており、説得を目的としたやりとりの中での参加者のパフォーマンスが総合的に評価される。

### 2.2.3 評価者

今回の分析に使用した 65 名の評価者の主要属性は表 2 のとおりである。なお、これらは、すべて、自己申告に基づく。

表 2

ICNALE GRA 発話評価者 65 名の主な属性プロフィール

| 属性          | 属性下位区分と人数                                                                                                                                                 |
|-------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| 性別          | 女性(32), 男性 (33)                                                                                                                                           |
| 年齢          | 20 代(16), 30 代(28), 40 代(17), 50 代(1), 60 代(3)                                                                                                            |
| 学位          | 高卒(1), BA(21), MA(28), Ph. D. (15)                                                                                                                        |
| 専攻分野        | 理工(6), 英語系(28)、人文(5), 医療(3), 英語以外の言語(3), 社会科学(20)                                                                                                         |
| 英語力(自己診断)   | B1(2), B2 (17), C1 (24), C2 (14), ENS または ENS 同等 (8)                                                                                                      |
| 職業          | ビジネス (8), 英語教員・翻訳者(33), 英語以外の教員(8) 無職(院生含)(7), その他(9)                                                                                                     |
| 過去のスピーチ評価経験 | 未経験(9), 1 回以上 (19), 6 回以上 (37)                                                                                                                            |
| 母語          | アラビア語(1), 中国語(10), 英語(2), フィリピン語(15), モン語(1), インドネシア語(3), 日本語(9), コンカニ語(1), 朝鮮語(3), ラオ語(8), マレー語(1), パンジャブ・ウルドゥ語(1), タイ語(6), ウルドゥ語(2), ウイグル語(1), ベトナム語(1) |

注: 専攻分野区分に関して、人文系のうち、英語系のみ独立した区分としている。

表 2 に明らかなように、ICNALE GRA では、母語話者の英語教師に限らず、多様な L1 背景や職業背景を持つ評価者を集めている。これにより、教室環境に限定されない、実社会における英語としての通用性・妥当性の評価が可能になる。

### 2.2.4 評価の手順

ICNALE GRA の評価は、全体的評価スコア(Holistic Assessment: HOL、100 点満点)と、10 観点別の分析的評価スコア(10 点満点×10=100 点)の両面で行われる。評価者はまず、対象の音声ファイルを聞いて、全体の印象に基づき、全体的評価スコアを決めた。その後、ICNALE 評価者用ガイドブックに示された評価観点(ループリック)をふまえ、分析的評価スコアを決定した。ループリックの正文は英語であるが(Ishikawa, 2023 に全体を掲載)、表 3 は、それらを抄訳したものである。

表 3

ICNALE 分析的評価の各観点

| タイプ | 観点                              | 評価のポイント                                   |
|-----|---------------------------------|-------------------------------------------|
| 言語系 | 理解可能度<br>(Intelligibility: INT) | 言語的なレベルで受信(読んで理解、聞いて理解)が可能か。内容的妥当性とは区別する。 |

|                        |                                                   |                                                                                         |
|------------------------|---------------------------------------------------|-----------------------------------------------------------------------------------------|
| 内 容 系<br>(Contents: C) | 複雑性<br>(Complexity: CLX)                          | 複雑な単語、フレーズ、表現、構文、文法を含むか。                                                                |
|                        | 正確性<br>(Accuracy: ACC)                            | 語彙や文法の面で間違いがないか。局所的な小さいミスは不問。母語話者ではなく、熟達した English as a lingua franca (ELF) 話者を基準に判断する。 |
|                        | 流暢性<br>(Fluency: FLU)                             | 量的に流暢か。流暢性を阻害する特徴(フィラーなど)が目立っていないか。                                                     |
|                        | 内容的妥当性<br>( Comprehensibility: CPH)               | 内容的に妥当で筋の通った考えが述べられているか。言語面での理解可能性とは区別される。                                              |
|                        | 論理性<br>(Logicity: LGC)                            | 主張の内容が、どの程度論理的で妥当か。理由と結論が論理的に接合しているか。                                                   |
| 態 度 系<br>(Attitude: A) | 洗練性<br>(Sophistication: SPH)                      | 提示されたアイデアは、洗練され、批判的に考え抜かれ、ユニークで、独創的で、革新的か。                                              |
|                        | 目的性<br>(Purposefulness: PPS)                      | タスクの目的を理解し、一貫して、その目的の遂行に注意を払っているか。                                                      |
|                        | コミュニケーション意欲<br>( Willingness to communicate: WIL) | コミュニケーションに意欲的か。発話中の発話量やターンテイクの頻度、作文中の文章量、提示されたアイデアの数、強意副詞使用頻度などが関係しうる。                  |
|                        | 対人関与度<br>(Involvement: INV)                       | 一方的に話したり書いたりするのではなく、自分をアピールして、聞き手や読み手を巻き込んだり、談話に参加させようとする姿勢があるか。                        |

ICNALE GRA では、全体的評価スコアおよび 10 観点別の分析的評価スコアに加え、観点別スコアを合算した分析的評価スコア合計 (Analytic Score Sum: ANAS) と、HOL および ANAS を平均化した統合評価指標 (ORS) が算出されており、評価指標は全体で 13 種となる。

なお、評価データの収集にあたっては、評価者における厳格性レベルのずれや、評点の過度の集中やばらつきを抑制するため、すべての評価者に、ルーブリックの精読と、ルーブリック内容の理解度を問うオンラインの修了テストの合格を義務付けた。また、評点の平均点を 40～60% (全体的評価は 45～55%)、評点の標準偏差を 20～30% の範囲に収めることを要請し、規準に反している場合はすべて修正を依頼した。また、質の高い評価データを収集するため、作業に必要となる総時間を積算の上で謝礼を支払っている。

## 2.3 分析手法

今回の分析では、140 名の発話サンプルの評点を比較するため、事前に、発話サンプルごとに 65 名の評価者による評価スコアを平均化しておく。

RQ1 については、13 種のスコアを変数、65 人の評価者をケースとする分割表に対して変数クラスター分析を行い、樹形図によって変数間の関係性を概観する。なお、1 次距離は  $2 \cdot 2r$  で定義

し、融合後の距離計算には Ward 法を使用する。

RQ2 については、10 観点別の分析的評価スコアの特徴の違いを加味した評点合成を行った場合の観点ごとのウェイトを観察するため、変数合成手法として、主成分分析と因子分析の 2 種を実施する。主成分分析で得られる第 1 主成分の主成分負荷量、因子分析で得られる第 1 因子の因子負荷量を概観した後、主成分得点、因子得点によるスコアデータの合成を行い、順位の変化を確認する。因子分析の共通性の初期値は SMC、因子推定法は最尤法、反復回数上限は 50、線形結合による除外はなし、回転は斜交プロマックス(次数 3) (ただし 1 因子の場合は回転しない)、規準化あり、とする。

RQ3 については、RQ1 と同じ頻度表に対して対応分析を実施し、第 1 次元・第 2 次元をそれぞれ横軸・縦軸とする散布図で、発話サンプルとスコアタイプとの関係を概観する。その後、散布図内の 4 象限の特性を 4 つの発話スタイルとみなして解釈し、各々を典型的に代表する発話サンプルの特定を目指す。

なお、今回の分析で使用した評価データの概要は表 4 の通りである。

表 4

評価データの概要

| 変 数   | サンプル数 | 平均     | 不偏分散    | 標準偏差   | 最小値    | 最大値    |
|-------|-------|--------|---------|--------|--------|--------|
| L_INT | 140   | 5.345  | 2.126   | 1.458  | 1.723  | 8.577  |
| L_CLX | 140   | 4.838  | 2.241   | 1.497  | 1.362  | 8.154  |
| L_ACC | 140   | 5.035  | 2.118   | 1.455  | 1.646  | 8.462  |
| L_FLU | 140   | 5.048  | 2.593   | 1.610  | 1.462  | 9.000  |
| C_CPH | 140   | 5.390  | 2.015   | 1.419  | 1.677  | 8.562  |
| C_LGC | 140   | 5.229  | 2.082   | 1.443  | 1.392  | 8.338  |
| C_SPH | 140   | 4.953  | 2.187   | 1.479  | 1.315  | 8.338  |
| C_PPS | 140   | 5.366  | 2.053   | 1.433  | 1.615  | 8.369  |
| A_WIL | 140   | 5.393  | 2.112   | 1.453  | 1.400  | 8.400  |
| A_INV | 140   | 5.198  | 2.031   | 1.425  | 1.385  | 8.508  |
| HOL   | 140   | 52.042 | 227.672 | 15.089 | 14.415 | 84.246 |
| ANAS  | 140   | 51.794 | 210.479 | 14.508 | 14.977 | 82.931 |
| ORS   | 140   | 51.918 | 218.862 | 14.794 | 14.696 | 83.588 |

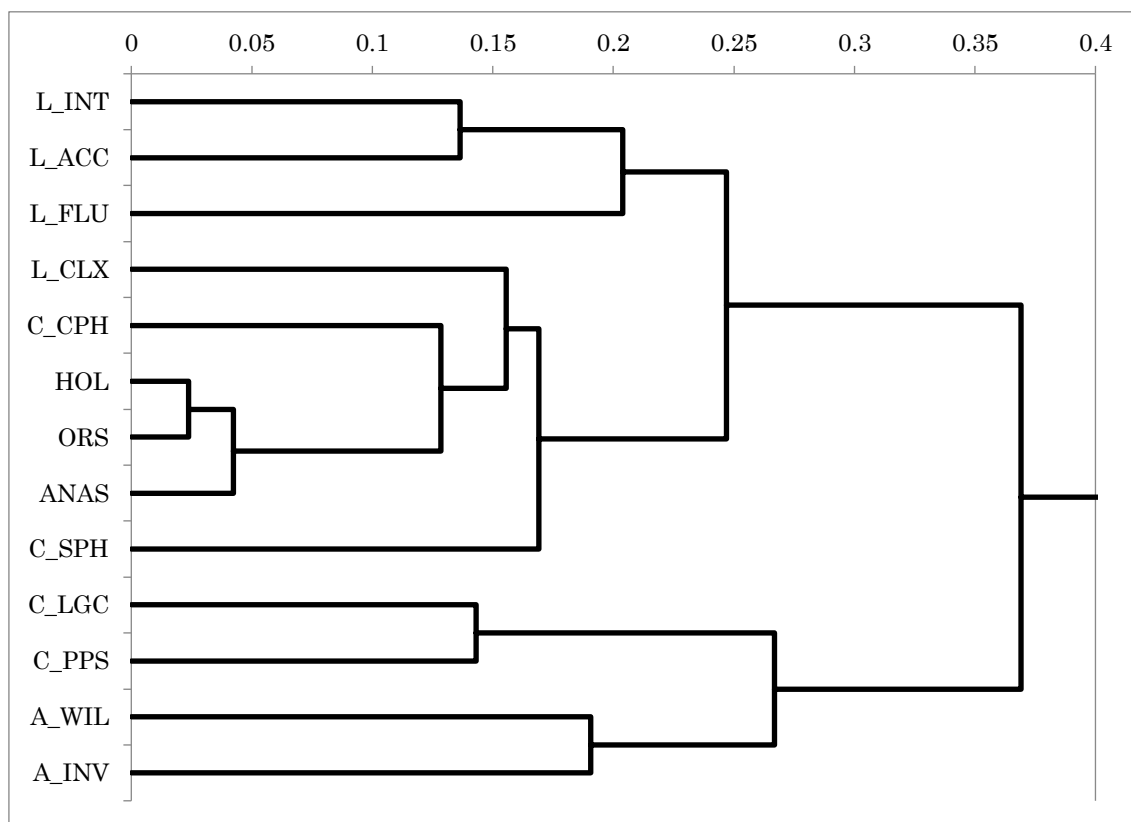
### 3. 結果と考察

#### 3.1 RQ1 指標間の関係

10 観点別の分析的評価スコア、および、3 種の統合系指標の相互関係を探るため、変数クラスター分析を実施したところ、図 1 の樹形図が得られた。

図 1

13 指標間の近接性(クラスター分析に基づく樹形図)



樹形図の解釈にあたっては、定常状態が最も長く続く 0.3 前後にカッティングポイントを置くのが妥当であろう。そのようにした場合、全体は、上下の二大クラスターに分割された。このうち、上部クラスターには、発話の言語面に関する理解可能性・正確性・流暢性・複雑性や、3 種の統合系指標が含まれる。一方、下部クラスターには、発話の内容に関する論理性と目的性、発話態度と関係の深いコミュニケーション意欲と対人関与の 4 観点が含まれる。

以上より、説得型ロールプレイの評価者は、(a)聞き手に理解可能で、文法的な正確性と複雑性を有する発話を流暢に産出することと、(b)説得というタスクの目的を明確に意識し、それに沿った論理的な内容を組み立て、かつ、相手を意識して積極的に主張を展開することの二面を評価においてとくに重視していたことがわかる。3 つの統合的指標が上部クラスターに入っていることから、評価の力点は(b)よりも(a)のほうに置かれている。

なお、ここで注目すべきことが 2 つある。1 点目は、大半の内容系観点が下部クラスターに入っている中で、内容上の理解可能性のみが上部クラスターに分類されたことである。ループリックを用意する際には、言語的な理解可能性(言っていることが聞き取れる)と、内容的な理解可能性(言っていることが納得できる)をあえて区別したわけだが、実際の評価では、これらはともに言語的な要素として扱われたようである。2 点目は、3 種の統合系指標の中で、10 観点の積み上げによる分析

的評価スコア合計 (ANAS) よりも、全体を総括的に評価する全体的評価スコア (HOL) のほうが、統合評価指標 (ORS) に早く融合していることである。今回のデータについて言えば、全体的評価のほうが最終評価に近い分布になっていると推察される。

### 3.2 RQ2 指標合成と指標別のウェイト

次に、10 観点別の分析的評価スコアのウェイトについて検証したい。前述のように、Ishikawa(2023)の処理では、観点別スコアの単純合算によって、つまりは、10 観点を、均等な重みを持つ相互に独立した変数とみなしてスコアの合成を行ったわけだが、仮に、多変量解析手法を用いたスコア合成において、線形結合が確認されたり、観点ごとに大きく異なるウェイトが与えられたりするようであれば、Ishikawa(2023)の処理には問題があったことになる。

そこで、主成分分析と因子分析を実施したところ、線形結合した変数は認められなかった。また、第 1 主成分の寄与率は 97.75%、第 1 因子の寄与率は 97.47%で、いずれも全変数の負荷量がプラスとなっており、10 観点別スコアを集約した総合指標が取り出せていることが確認された。各変数にかかるウェイトは、表 5 に示すとおりである。

表 5

第 1 主成分および第 1 因子において、各変数に与えられるウェイト

| 変数    | 主成分負荷量 | 因子パタン  |
|-------|--------|--------|
| L_INT | 0.9916 | 0.9925 |
| L_CLX | 0.9928 | 0.9929 |
| L_ACC | 0.9857 | 0.9867 |
| L_FLU | 0.9847 | 0.9830 |
| C_CPH | 0.9946 | 0.9954 |
| C_LGC | 0.9888 | 0.9874 |
| C_SPH | 0.9922 | 0.9914 |
| C_PPS | 0.9869 | 0.9834 |
| A_WIL | 0.9863 | 0.9817 |
| A_INV | 0.9829 | 0.9781 |

主成分負荷量、因子パタンともに、ウェイトはすべての変数でほぼ 1 に近く、極端にウェイトが大きい、あるいは、小さい変数は確認されなかった。このことより、10 観点別の分析的評価スコアの単純合算が、統計的に、大きな問題を含むものではなかったことが確認された。ICNALE GRA のプロジェクトでは、ループリックの内容理解テストを課す形で評価者のトレーニングを行っており、加えて、全評価者に対して、スコアの平均値と標準偏差を指定範囲内に収めるよう要求している。このため、観点ごとの評価データのばらつきが抑制されたものと推測される。

なお、大きな差ではないが、負荷量を降順で並べた場合、言語系の指標が上位に、内容系の指標が次位に、態度系の指標が下位にかたまる。これは、内容・態度系指標に比べ、言語系の指標はより評価しやすく、評価結果が安定しやすいことを示唆する。この結果は、図 1 で見たように、3 種の統合系指標が、内容・態度指標よりも言語系指標と近い関係にあることと符合する。

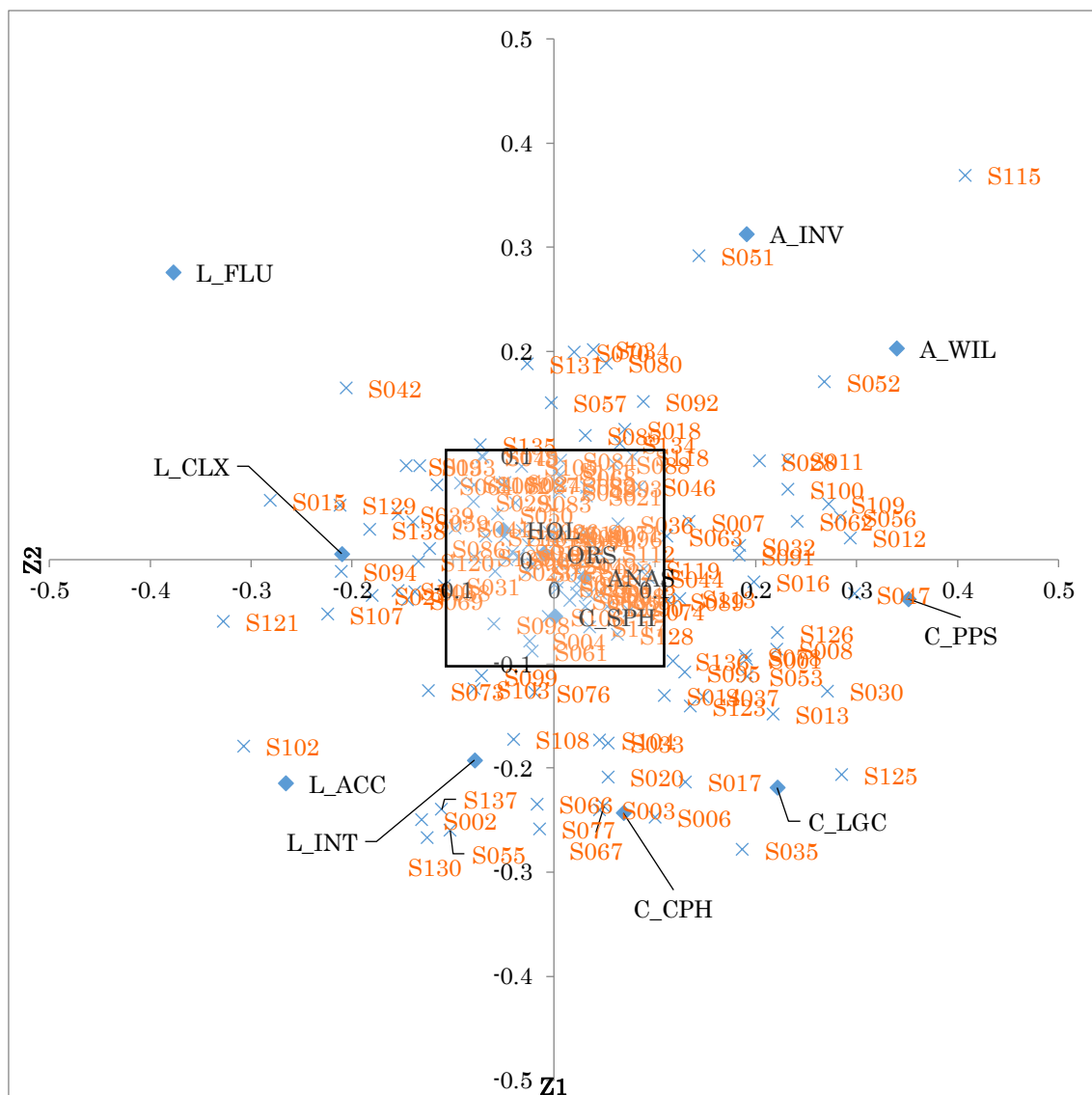
### 3.3 RQ3 発話スタイル別ベンチマークサンプルの特定

最後に、13 種の評価指標と各発話サンプルの関係を可視化し、主要な発話スタイルとそれらを典型的に表す発話サンプルを探るため、対応分析を実施した。

抽出された次元 1 の説明力は 40.3%、次元 2 の説明力は 18.3%で、上位 2 次元で元の分散の過半が説明されることが確認された。そこで、次元 1 を横軸、次元 2 を縦軸として散布図を描画したところ、図 2 の結果を得た。なお、下図中の S001～S140 は、140 本の発話サンプルに対してランダムに付与された連番コードであり、連番の近さは話者属性の近接性を示すものではない。

図 2

13 指標と 140 本の発話サンプルの関係性(対応分析に基づく散布図)



はじめに、2つの次元の持つ意味について概観したい。次元1(横軸)では、言語系(L\_)の指標がすべて左側にかたまっており、右側には、態度系(A\_)と内容系(C\_)の指標がかたまっている。このことから、次元1は、発話の言語特性と非言語特性を峻別する軸であると解釈できる。一方、次元2(縦軸)では、態度関連の2指標(コミュニケーション意欲・対人性)と流暢性の3指標が上部にかたまり、下側には、その他の言語系・内容系指標がかたまっている。このことから、次元2は意欲的・対話的で能弁な発話と、より抑制的な発話、言い換えれば、情意的な発話と理知的な発話を峻別する軸であると解釈できる。

これらの2つの次元によって座標空間は4つの象限に分割される。各象限を構成する両軸上の特性、および、そこに布置された評価指標の特性をふまえると、4つの象限は、それぞれ、異なる発話スタイルを表していることになる。以下、右上の象限を第1象限、時計回りの逆順で各象限を第2、第3、第4象限と呼ぶ。

まず、第1象限は情意的・非言語的な領域で、語彙や文法といった言語面を重視するよりも、対話の相手を強く意識し(INV)、相手とのリアルタイムのやりとりを意欲的に進めようとする(WIL)発話スタイルを表す。第2象限は情意的・言語的な領域で、相手に対してできるだけ多く話そうとする(FLU)と同時に、言語的な複雑性(CLX)にも一定の配慮を払った発話スタイルである。第3象限は理知的・言語的な領域で、発話量を増やしたり相手と激しいやりとりをしたりするよりも、聞き取りやすく(INT)、文法的に正確な(ACC)産出を行うことを優先する発話スタイルである。第4象限は理知的・非言語的な領域で、言語的な質以上に、発話内容の質、つまりは、聞き手を納得させるに足る、妥当で(CPH)、論理的で(LGC)、目的に沿った(PPS)産出を行うことを優先する発話スタイルである。

さて、各象限には各種の評価指標、また、発話サンプルが布置されているわけだが、原点近傍に位置するものは、特定の象限特徴を強く示すというよりも、むしろ中立的な性質を持っていると考えた方が妥当であろう。仮に、原点を中心とする $\pm 0.1$ の範囲を近傍部とみなすと(図2の中央部の黒線で囲まれた領域)、13種の評価指標のうち、3種の統合系指標(HOL, ORS, ANAS)と、内容系の洗練性指標(SPH)が原点近傍部に入った。説得型のロールプレイにおいて、発話の内容が十分に練られたものになっているかどうかは、特定の発話スタイルに関するものではなく、統合系指標と同様、それらを超えた基盤的な要素であると言えそうである。

一方、発話については、全140種のサンプルのうち、60サンプルが原点近傍部に入った。これらを除く80サンプルは、いずれかの象限特徴を持つと言えるが、象限特徴を2つの次元の特徴の組み合わせで定義していることを考えると、いずれかの次元において原点近傍に位置するサンプルの象限特徴は相対的に弱いと見るべきであろう。そこで、原点近傍の概念を拡張し、次元1、次元2の一方が $\pm 0.1$ の範囲に入っているサンプルを探すと、さらに59サンプルが該当した。以上の処理で残った21サンプルは、いずれの次元においても、 $\pm 0.1$ の範囲外にあり、当該象限の特性を顕著に反映している発話と考えられる。本研究では、これら21サンプルを、4種の発話スタイル別のベンチマークサンプルとみなす。

以上を整理したのが表6である。表6においては、解釈の参考のため、発話サンプルコードの

後に、各学習者の地域(CHN:中国、IDN:インドネシア、JPN:日本、KOR:韓国、THA:タイ、TWN:台湾)と習熟度(初級:A2、中級下:B1\_1、中級中:B1\_2、上級以上:B2+)を示すコードを追加で記載している。紙幅の関係で、個々のサンプルの実例を示すことは避けるが、すべてのサンプルは ICNALE のウェブサイトからダウンロードが可能である。

表 6

4 種の発話スタイルと、発話スタイル別ベンチマークサンプル

| 象限 1<br>(発話スタイル 1)                                    | 象限 2<br>(発話スタイル 2)                                     | 象限 3<br>(発話スタイル 3)                                                                                         | 象限 4<br>(発話スタイル 4)                                                                                                                                                                                      |
|-------------------------------------------------------|--------------------------------------------------------|------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 【情意・非言語】 対話の相手を強く意識し、相手とのリアルタイムのやりとりを意欲的に進めようとする(3 種) | 【情意・言語】 相手に対してできるだけ多く話そうとすると同時に、言語的な複雑性にも一定の配慮を払う(1 種) | 【理知・言語】 聞き取りやすく、文法的に正確な産出を行うことを優先する(6 種)                                                                   | 【理知・非言語】 聞き手を納得させるに足る、妥当で、論理的で、目的に沿った産出を行う(11 種)                                                                                                                                                        |
| S051 (KOR_B1_1)<br>S052 (THA_B1_1)<br>S115 (TWN_A2)   | S042 (IDN_B2+)                                         | S073 (IDN_B12)<br>S002 (KOR_B2+)<br>S130 (KOR_B1_1)<br>S102 (KOR_B1_2)<br>S055 (TWN_A2)<br>S137 (TWN_B1_2) | S053 (CHN_B1_1)<br>S095 (CHN_B1_1)<br>S125 (JPN_A2)<br>S037 (JPN_A2)<br>S017 (JPN_B1_1)<br>S030 (JPN_B1_1)<br>S123 (JPN_B1_1)<br>S006 (JPN_B1_2)<br>S014 (JPN_B1_2)<br>S035 (THA_A2)<br>S013 (TWN_B1_1) |

抽出されたベンチマークサンプルについて注目すべきこととして、ここでは 3 点を指摘しておきたい。

まず、1 点目は、母語話者サンプルと ESL 圏 4 地域の学習者サンプル合計 24 本のうち、ベンチマークサンプルに入ったものが皆無であり、また、B2+レベルの EFL 圏学習者サンプル合計 30 本のうち、ベンチマークサンプルに入ったものが 2 本(S042 と S002)のみだったことである。このことは、特定の発話スタイルが強く出やすいのは初中級の EFL 圏学習者の発話で、習熟度が初中級から上級になるにつれて、また、地域が EFL 圏から ESL 圏に移行するにつれて、さらには学習者から母語話者に変化するにつれて、より一般的・中立的な発話スタイルになっていくことを示す。

2 点目は、4 種の発話スタイルのうち、理知的な言語産出に関わる発話スタイル 3、4 に該当するサンプルが多いのに対し、情意的な言語産出に関わる発話スタイル 1,2 に該当するサンプルが少ないことである。このことは、EFL 圏初中級学習者のロールプレイ発話において、情意面、つまりは、相手を説得する強い意思や意欲の感じられる産出が少なかったことを示す。

3 点目は、一部の発話スタイルが特定地域の学習者と強く結びついていることである。EFL の 6

地域からはそれぞれ同数のサンプルを取っているわけだが、発話スタイル 3 に合致する 6 サンプルの半数は韓国の学習者のもので、発話スタイル 4 に合致する 11 サンプル中、7 サンプルは日本の学習者のものであった。今回のデータに限って言えば、同じ初中級レベルであっても、韓国の学習者発話は内容面よりも言語面に、日本の学習者発話は言語面よりも内容面に強みがある(言い換えれば、その逆が弱みとなっている)可能性がある。

こうして、Ishikawa (2023)で提案したレベル別のベンチマークサンプルを補完するものとして、4 種の発話スタイル別のベンチマークサンプルを抽出することができた。これらは、個人ないし数名の判断によるものではなく、65 名の評価者の評価データから統計的に取り出されたものであるという点で、レベル別ベンチマークサンプルと同様の信頼性を遡する。レベルサンプルとスタイルサンプルを組み合わせることで、アジア圏学習者の英語発話の特徴をより多面的に分析することが可能になるだろう。

#### 4. まとめ

以上、本研究では、ICNALE GRA の発話評価データを再分析し、評価指標間の相互関係(RQ1)、分析的評価スコアの合成に多変量解析手法を用いた場合の観点別ウェイト(RQ2)、発話スタイルと発話スタイル別ベンチマークサンプルの抽出(RQ3)を行った。

RQ1 については、13 種の評価指標が 2 群に大別されること、言語面(聞き手に理解可能で、文法的な正確性と複雑性を有する発話を流暢に産出すること)と、内容・態度面(説得というタスクの目的を明確に意識し、それに沿った論理的な内容を組み立て、かつ、相手を意識して積極的に主張を展開すること)が説得型ロールプレイにおける良い発話を評価する際の二大基準になっていること、発話の全体的な質は主として前者によって決定すること、などが明らかになった。

RQ2 については、10 観点別の分析的評価スコアは 1 個の代表値に集約が可能であること、その際、各観点にかかるウェイトはほぼ同等であることが確認された。これにより、10 観点別の分析的評価スコアの合算値を統合評価指標(ORS)に組み込んだ Ishikawa (2023)の処理が結果的に妥当なものであったことが傍証された。

最後に、RQ3 については、対応分析により、「対話の相手を強く意識し、相手とのリアルタイムのやりとりを意欲的に進めようとする」、「相手に対してできるだけ多く話そうとすると同時に、言語的な複雑性にも一定の配慮を払う」、「聞き取りやすく、文法的に正確な産出を行うことを優先する」、「聞き手を納得させるに足る、妥当で、論理的で、目的に沿った産出を行う」という 4 種の発話スタイルの存在が示唆され、各々のスタイルについて、その特性を最も顕著に示す 21 種のサンプルを特定することができた。また、個々の発話スタイルの特徴は EFL 圏の初中級学習者に出やすいこと、また、同じ EFL 圏でも国によって差がある可能性が示唆された。

本稿執筆時点において、ICNALE GRA は最終段階のデータ収集を継続しており、2022 年度末には、合計 150 名の評価者(発話評価者・作文評価者各 75 名)による評価データセットとしてリリースされる予定である。アジア圏学習者の L2 産出に対するユニークな大型評価データセットが、今後、広く研究や教育の場で使用されることを希望したい。

## 謝辞

本研究は科学研究費プロジェクト「アジア圏英語学習者の作文・発話に対する世界最大級の公開型評価データセット開発と分析」(20H01282)による研究成果の一部である。

## 引用文献

- Ishikawa, S. (2019). The ICNALE Spoken Dialogue: A new dataset for the study of Asian learners' performance in L2 English interviews. *English Teaching* (The Korea Association of Teachers of English), 74(4), 153-177.
- Ishikawa, S. (2020). Aim of the ICNALE GRA project: Global collaboration to collect ratings of Asian learners' L2 English essays and speeches from an ELF perspective. *Learner Corpus Studies in Asia and the World*, 5, 121-144.
- Ishikawa, S. (2023). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.
- Ishikawa, S. (in press-a). Effects of raters' L1s, assessment experiences, and teaching experiences on their assessment of L2 English speeches: A study based on the ICNALE Global Rating Archives. *Learn Journal: Language Education and Acquisition Research Network*.
- Ishikawa, S. (in press-b). A new yardstick of comparison for contrastive interlanguage analysis: A study based on the ICNALE Global Rating Archives. *Proceedings of the Asia TEFL 2022 Conference*.