日本語学習者の作文に対する安定した印象評価データの収集方法

本多由美子 [a]　　井伊菜穂子 [b]

[a] 国立国語研究所 研究系

[b] 琉球大学／国立国語研究所 共同研究員

責任著者：本多由美子（honda@ninjal.ac.jp）

要旨

本稿では，日本語学習者縦断作文コーパス『W-CoLeJa』に付与する予定の印象評価データについて，安定したデータの収集方法を検討した。日本語母語話者 12 名が 1 人当たり 42 作文を評価したパイロット調査の結果を，相関分析および一般化可能性理論によって分析したところ，3 点の結果が得られた。（1）中程度の作文を評価の目安とすることで評価者間の評価尺度のずれを小さくし，ある程度安定したデータを収集することができる。（2）収集方法の改善点として，総合評価を 50 点満点の細かい評価から 10 段階などの粗い評価にする，一度に評価する作文数を減らす，説明の中の「結束性」のような専門的な語を平易な語に置き換えることが必要である。（3）評価者数について，一般化可能性理論の D 研究の結果から，評価項目数が現行の 11 項目の場合，現行の調査より評価者数を減らしてもある程度の信頼性を確保することが可能である*。

キーワード：『W-CoLeJa』，ライティング，日本語母語話者，クラウドソーシングサービス，一般化可能性理論

1. 背景

「よく書けている」と感じる作文とはどのような作文であろうか。国立国語研究所共同研究プロジェクト「日本語学習者の作文の縦断コーパス研究」（以下，「本プロジェクト」とする）では，この問いを明らかにすることを目標に，クラウドソーシングサービスを通じて一般の日本語母語話者から日本語学習者が執筆した作文に対する印象評価のデータを収集している。本稿は，本調査に先立って行ったパイロット調査の結果を用い，安定した印象評価データを得るための収集方法を検討することを目的とする。なお，評価対象の作文は，現在構築中の日本語学習者縦断作文コーパス『W-CoLeJa』に収録予定の作文である。また，本プロジェクトで収集する印象評価

*本稿は日本語学会 2024 年度春季大会（2024 年 6 月 2 日於東京外国語大学）のワークショップでの発表内容をもとに作成したものである。ワークショップでコメントをくださったみなさま，本稿の執筆に際して貴重なご意見をくださった水本篤先生（関西大学），本稿の掲載にあたり貴重なご指摘，ご助言をくださった査読委員の先生に厚く御礼申し上げる。また，本稿は国立国語研究所の共同研究プロジェクト「多様な言語資源に基づく日本語非母語話者の言語運用の応用的研究」（共同研究番号：L411062227）のサブプロジェクト「日本語学習者の作文の縦断コーパス研究」（いずれもプロジェクトリーダー：石黒圭），JSPS 科研費 20K02974，21H04417 および 23K02509 の研究成果である。本研究の遂行にあたり，国立国語研究所の上記プロジェクトのスタッフ各位および科研の関係者，調査協力校の教員・学習者のみなさまのご協力を得た。記して感謝申し上げる。

データは，将来的に作文の自動添削・自動評価システムなどの構築に活用することを視野に入れている。

日本語教育のライティング評価研究で人による評価を扱った従来の研究では，主に日本語教師や大学教員を評価者としている（田中・長阪 2006，田中 2016 など）。アカデミック・ライティング教育支援のためには，その評価の実態を明らかにすることは重要である。

しかし，「おもしろい内容だ」や「共感できる文章だ」のような読み手が作文から受ける印象も評価に含めるとすれば，日本語学習者（以下，学習者とする）が自らの文章を評価される場面は教育の場に限らない。ビジネス文書や就職活動のエントリーシート，知り合いに近況を伝えるメールなど，学習者の書いた文章が教育現場の外で教師以外の日本語母語話者から評価を受ける場面もあるだろう。一般の日本語母語話者による評価を対象とした研究は田中他（1998）や宇佐美（2014）などに限られ，その実態は十分に把握されてはいない。

ライティングのようなパフォーマンスに対する評価では，評価者間の評価が一致しにくいことが課題として指摘されている（宇佐美 2014，田中 2016）。本プロジェクトにおいて測ろうとする，作文を読んだ後の印象は，各評価者の感じ方の違いが見られることが予想され評価を一致させることは難しい面もあると思われる。この点について本プロジェクトでは，クラウドソーシングサービスを通じて 1 本の作文に対し複数名から評価データを収集することによって，より安定したデータを得ることを計画している。また，複数名から評価を集めたときに，学習者の作文の中には評価が一定の範囲内に収まるものと，ばらけるものがあることも予想される。このような評価者の印象の実態を把握するには，評価作業において，評価の尺度はできるだけ揃える必要がある。そこで本稿では，印象評価調査の本調査の前に行ったパイロット調査の結果を用い，評価作業において評価者間における評価尺度のずれを小さくするような作業の方法について検討する。

2. 先行研究

日本語教育のライティング評価研究において，人による評価を扱った従来の研究は，主に日本語教師や大学教員を評価者としている。この分野の研究には，田中真理氏を中心とする「Good Writing」の一連の研究（田中・長阪 2006，2009，田中 2016，2022 など）や伊集院（2017），伊集院他（2020）などがある。

人がライティングのようなパフォーマンスを評価する場面では，評価者間の評価の不一致という問題がある（田中・長阪 2009，田中 2016）。試験の採点や授業の評定を行うときなど，客観性や公平性が求められる場面において，評価の不一致は大きな問題である。そのため，評価者間の評価の不一致をできるだけ減らすために，評価基準の策定や評価用のフローチャートの開発（田中・長阪 2006，2009，田中 2022）が行われている。

一方，日本語教師ではない，一般の日本語母語話者による評価を対象とした研究は田中他（1998），宇佐美（2014）などに限られる。田中他（1998）の調査は，作文の評価において日本語教師ではない一般の日本語母語話者による評価と日本語教師による評価を比較する目的で行われ，両者の重視する点には相違があること，一般の日本語母語話者よりも日本語教師のほうが評価にばらつ

きが少ないことが指摘されている。また宇佐美（2014）では，日本語学習者の書いた謝罪文に対する一般の日本語母語話者の評価を分析し，日本語母語話者が評価を行う際の評価の多様性を「評価プロセスモデル」を作成することで可視化している。

以上のように，先行研究からは一般の日本語母語話者によるライティング評価が多様であることは明らかになっているが，それを踏まえた上でどうすれば安定した評価データを収集できるかという視点での検討は十分に行われていない。

3. 研究課題

本プロジェクトでは，コーパス活用のためのメタ情報として，ある程度まとまった本数の作文に印象評価のデータを付与することを計画している。そのため，評価作業は安定したデータを収集することが重要となる。評価者によって作文から受ける印象には違いがあることを認めながらも，評価者間の評価の尺度のずれは小さくすることを目指す。なお，本稿での「安定したデータを収集する」とは同じ作文であれば似たようなデータが得られることを指す。本プロジェクトでは評価者間の尺度のずれを小さくしても，作文の中には評価が一定の範囲内に収まる作文と評価がばらける作文があることを想定しており，それも含めて似たようなデータが得られることを目標とする。

田中（2022）では評価の不一致をもたらす三つの要因に「A.スクリプト（作文）」「B.評価基準や評価方法」「C.評価者」が挙げられているが，本稿ではその中の「B.評価基準や評価方法」および「C.評価者」に注目する。まず，「A.スクリプト（作文）」については，先述のように評価が一定の範囲内に収まる作文と評価がばらける作文の特徴を明らかにすることが，本プロジェクトで今後取り組むテーマの一つであり，本稿ではそのための調査方法を検討している。次に「B.評価基準や評価方法」は，本稿で検討できる余地の最も大きい要因である。そして「C.評価者」は，特に後述する項目別評価においては評価者間で作文に対する感じ方に違いが見られることが予想され，すべての評価を一致させることは難しいと思われる。そこで本プロジェクトでは，クラウドソーシングサービスを通じて 1 本の作文に対し複数名から評価データを収集することによってより安定したデータを得る。1 作文当たりの評価者は多いほうがよいと思われるが，限られた資源で多くの作文に評価をつけることも求められる。そのため，どのくらいの人数の評価者がいれば安定したデータが得られるかも検討する必要がある。このように，本稿では「B.評価基準や評価方法」および「C.評価者」の人数を検討し，その上で本調査を実施することにより，より安定した評価データを収集できるものと考える。

以上のことから，本稿では本調査に先立って行ったパイロット調査の結果をもとに以下 3 点の研究課題（以下，RQ と略す）を設定する。

RQ.1　パイロット調査で安定した印象評価のデータが得られたか。

RQ.2　安定したデータを得るために調査方法をどのように改善すればよいか。

RQ.3　１作文当たり何人ぐらいの評価者に評価を依頼すれば，安定したデータが得られるか。

なお，これ以降，本稿の「評価」とは，作文に対する「印象評価」を指す。

4. 調査方法

ここでは，評価対象の作文の選定，評価項目，評価データの収集方法の 3 点について述べる。

4.1 評価対象の作文

評価の対象とした作文は，国立国語研究所共同研究プロジェクト「日本語学習者の作文の縦断コーパス研究」で収集中のデータから選んだ作文 42 本である。そこで，まずは本プロジェクトの概要を説明する。

本プロジェクトは，海外の大学で日本語を学び始めた学習者が大学 4 年間でどのように文章執筆能力を身につけていくのかを明らかにすることを目的としたプロジェクトである。調査協力校は，中国語（簡体字）圏 8 大学，中国語（繁体字）圏 2 大学，ベトナム語圏 3 大学，韓国語圏 2 大学，タイ語圏 2 大学，英語圏 1 大学，フランス語圏 1 大学，スロベニア語圏 1 大学の計 20 大学（2024 年 6 月時点）である。

作文のテーマには A テーマ群と B テーマ群があり，大学によって A テーマ群で依頼している大学と B テーマ群で依頼している大学がある。学習者は年に 3 回異なるテーマ（図 1 の①～③）で作文を執筆し，それを 4 年間繰り返し継続する（図 1）。このようなデータの収集方法により，特定の学年におけるテーマ間の比較を行うだけでなく，一つのテーマを対象に 4 年間の習得過程を追うことが可能となる。データの収集はすべてオンライン上で行っており，作文はキーボードで入力されている。なお，作文の文字数の目安は表 1 に示した通りである。作文を学年別に比較することを考えると 1 年次から 4 年次まで同程度の文字数であることが望ましいが、日本語を学び始めて日の浅い 1 年生に 1,000 字を超えるような長い作文を求めることは現実的に困難であること、一方で文章構造などの分析を行うにはある程度の長さの作文であることが望ましいことから、表 1 のような文字数の目安を設定した。

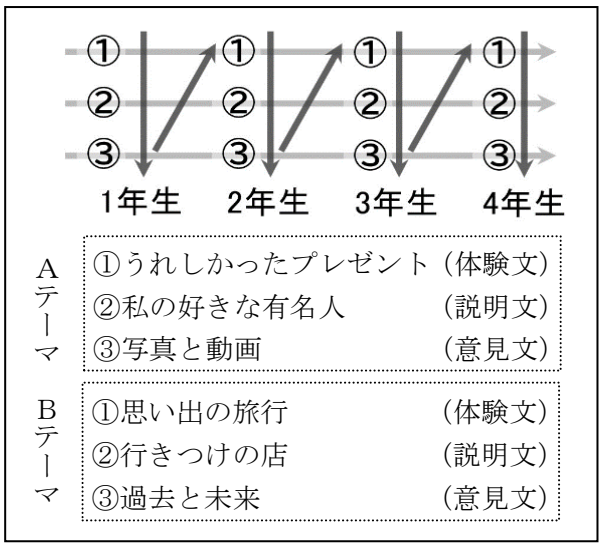

図 1　作文テーマとデータ収集の流れ

表 1　作文の文字数の下限と目安

学年	下限	目安
1 年生	200 字	400 字
2 年生	400 字	800 字
3 年生	600 字	1,200 字
4 年生	600 字	1,200 字

パイロット調査では，先述の方法で収集した A テーマ群の作文データの中から 42 本の作文を評価対象として選定した。その内訳を表 2 に示す。作文を選定する際は，各母語，各学年を代表する作文として，作文執筆時に示した字数の目安（前掲表 1 参照）に近い作文を選んだ。なお，評価対象の作文には日本語母語話者の作文も含まれるが，これは日本国内の大学の学部 1 年生を対象に、文字数の目安を 1,200 字とし，学習者と同様のテーマで別途収集したものである。

表 2　評価対象の作文（単位：本）

種類	テーマ	母語			計
		中国語	ベトナム語	日本語	
体験文	うれしかったプレゼント	1〜3 年 各 2	1〜3 年 各 2	2	14
説明文	私の好きな有名人	1〜3 年 各 2	1〜3 年 各 2	2	14
意見文	写真と動画	1〜3 年 各 2	1〜3 年 各 2	2	14
計		18	18	6	42

4.2　評価項目

本節では評価項目について述べる。なお，「4.2 評価項目」と「4.3 評価データの収集方法」は本プロジェクトで検討の上，決定したものである。

作文の評価は，総合評価，全体の理解[1]，項目別評価の 3 種類に分けて行った。総合評価は 1〜50 点（1 点刻み），全体の理解と項目別評価（表 3 の 11 項目）は 5 段階評価である。

表 3 に挙げた項目別評価の項目は，田中・長阪（2006）と石黒（2017）を参考に設定した。田中・長阪（2006）では，第 2 言語としての日本語ライティング評価の項目として「1.目的・内容」「2.構成・結束性」「3.読み手」「4.日本語（言語能力）」の 4 項目が挙げられ，それぞれの基準説明の欄に具体的な観点が述べられている。例えば，「2.構成・結束性」の基準説明には「文章全体の構成」「適切な段落分け」「段落と段落の関係」「段落内の文のつながり」がある。田中・長阪（2006）では 4 項目のみを評価項目としているが，本パイロット調査では基準説明に含まれているような具体的な観点もそれぞれ評価項目として設定した。また，ビジネス文書の分析を行った石黒（2017）では，クラウドソーシングサービスの発注文書における他者配慮について，情

[1] 「全体の理解」は，作文に書かれている事柄を理解した上で評価が行われたかを確認するために設けたものである。

報面の配慮とは別に「受注者（読み手）が文書を読んでどう感じるか」（石黒 2017: 31）という情緒面の配慮にも言及している。本パイロット調査でもこの点を参考にし，「3.情報面での配慮」と「4.心理面での配慮」という他者配慮に関わる二つの項目を取り入れた。

表 3　評価項目

項目		評価内容
総合評価（1〜50 点，1 点刻み）		作文を総合的に判断する。主観的な評価でよい。42 本の中で中程度の作文を「平均的（21~30 点）」として評価。
全体の理解（5 段階評価）		書かれている事柄がどのくらい理解できたか。5「すべてわかった」，1「まったくわからなかった」で評価。
項目別評価（11 項目 5 段階評価）		42 本の中で中程度の作文を「平均的（3）」として評価。
内容	1.発想力	独創的で読んでいておもしろい。
	2.説得力	読んでいてなるほどと思わされる。
	3.情報面での配慮	必要な情報が過不足なく書かれている。
	4.心理面での配慮	読んでいて不快感を覚えず温かい気持ちで読める。
	5.共感力	書き手や登場人物の気持ちが理解できるように書かれている。
構成	6.構成	文章全体がわかりやすい構成で書かれている。
	7.結束性	文と文とのつながりがなめらかで自然である。
	8.一貫性	文章の筋が通っていてまとまりがある。
表現	9.正確さ	表現が日本語として正しく使われている。
	10.ふさわしさ	作文の内容に合った自然な表現が選ばれている。
	11.豊かさ	さまざまな種類の表現が使われている。

上記の評価項目について，5 段階評価（総合評価は 1〜50 点の 1 点刻みの評価）をする際の目安を示したものが表 4 である。総合評価と項目別評価は評価する作文全体の中で中程度の作文を「平均的（3）」（総合評価では 21~30 点）として評価基準を定めた。これは，パイロット調査に先立って行われた予備調査で，評価者によって評価の厳しさに差があることが見受けられたためである。構築中の『W-CoLeJa』には初級から上級以上まで，幅広いレベルの作文が収められる。前掲の表 2 のように，レベルの広がりを意識して調査対象の作文を選び，その中の中程度の作文を 3 とすることで，評価者ごとの評価の尺度を調整しながら評価データが収集できるのではないかと考えられる。

表 4　評価基準

評価の目安	総合評価（50 点満点）	5 段階評価
平均より非常に優れている	41〜50 点	5
平均よりやや優れている	31〜40 点	4
平均的（42 本の中で中程度の作文）	21〜30 点	3
平均よりやや劣っている	11〜20 点	2
平均より非常に劣っている	1〜10 点	1

4.3 評価データの収集方法

評価データは，株式会社クラウドワークスが提供するクラウドソーシングサービス[2]を通して募集した評価者によるものである。評価者は，日本語母語話者 13 名である。13 名のうち評価作業の内容に不備が見られた 1 名を除き，本稿では 12 名を分析対象とした。12 名の性別の内訳は男女それぞれ6名であり，外国人が書いた文章を添削した経験を有する評価者が6名含まれている。

パイロット調査は次の手順で行った。まず，調査参加に関する説明書を渡し，同意を得た後，フェイスシートで年齢・性別・母語・出身地・職業[3]・日本語教育経験の有無・日本語非母語話者の文章の添削経験の有無を尋ねた。次に，学習者が書いた作文 42 本（4.1 参照）の PDF ファイルをまとめて渡し，前掲の表 3，表 4 の評価項目および基準に従ってエクセルファイルに総合評価（1～50 点），全体の理解（5 段階評価），項目別評価（5 段階評価）の評価結果を入力してもらった。なお，項目別評価については評価の理由の入力も依頼した。調査参加への同意とフェイスシートの回答は Google Form で，ファイルのやり取りはオンライン上で行った。

なお，本パイロット調査では，作文を形態素解析したときに誤解析が生じるレベルの誤用（表記の誤りや活用の誤りなど）を事前に修正した上で，評価対象の作文として用いている[4]。これは，学習者の作文を読み慣れない評価者が表記の誤りから生じる読みにくさによって，内容の評価が十分に行えないことへの配慮による。本プロジェクトでは，漢字使用などの表記が十分に身についていないレベルの学習者による作文であっても，内容面が優れているものがあると予想しているため，評価者には表記の誤りにとらわれすぎない評価を行ってもらいたいと考えている。しかし，修正箇所の多い作文を評価することは，日本語母語話者の評価の実態を探るという本プロジェクトの目的から外れる。そのため，これら 2 点の事情を合わせて，パイロット調査では形態素解析を行ったときに，誤解析が生じるレベルの誤りを訂正するに留めた[5]。

5.分析方法

RQ を明らかにするため，評価者 12 名が 42 作文に対して行った評価のうち，総合評価と項目別評価のデータを用いて分析を行う。分析に用いるデータと RQ との対応を図 2 にまとめた。

まず，総合評価（1～50 点，1 点刻み）のデータを用い，点数の分布や評価者間の相関をもとに，RQ.1「パイロット調査で安定した印象評価のデータが得られたか」を検討する（分析①）。そして，その検討の過程で RQ.2「安定したデータを得るために調査方法をどのように改善すればよいか」の示唆を得る（分析②）。次に，項目別評価（11 項目，5 段階評価）のデータを用い，「作文」「評価項目」「評価者」の三つの要因が評価値のばらつき（分散）にどの程度影響を与えて

---

[2] クラウドソーシングサービスとは，オンライン上で「仕事をしてほしい人（依頼者・クライアント）と，仕事をしたい人（受注者・ワーカー・メンバー）を，効率よく繋いでくれるサービスのこと」（クラウドワークスのホームページ（https://crowdworks.jp/articles/5926/ 最終確認日:2024 年 8 月 28 日）より）である。

[3] 評価者 12 名の職業の内訳は，会社員 4 名，主婦 2 名，介護士 2 名（福祉系と回答した方 1 名を含む），プログラマー1 名，医療専門職 1 名，自営業 1 名，フリーランス 1 名である。

[4] 事前修正が必要な箇所は実際には少なく，1 作文あたり平均 1.69 箇所（日本語母語話者の作文を除く）であった。

[5] 誤用の中には，語彙の選択や文体レベルの誤用など，形態素解析が適切になされる誤用もある。このような誤用については事前に修正を行っていない。

いるかを検討する（分析③）。これは RQ.1 と対応する。この検討には一般化可能性理論（generalizability theory）の G 研究（G study）を用いる。一般化可能性理論については後述する。この検討の過程で RQ.2 の示唆を得る（分析④）。そして，分析③の G 研究で算出された結果を用いて，RQ.3「1 作文当たり何人ぐらいの評価者に評価を依頼すれば，安定したデータが得られるか」を検討する（分析⑤）。この検討には一般化可能性理論の D 研究（D study）を用いる。

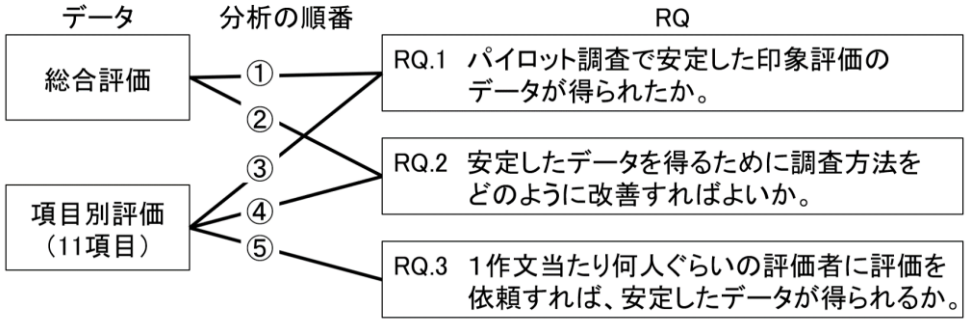

図 2　分析データと RQ との関係

6.結果

6.1 総合評価

まず，総合評価（1～50 点）を用いて，RQ.1 を検討する。図 3 は評価者 12 名の総合評価点の分布である。箱ひげ図に各作文の点数をプロットした。黒い三角形は平均点を表す。12 名の中で平均と中央値がいずれも最も低いのは評価者 6（平均 20.2，中央値 18），最も高いのは評価者 9（平均 31.6，中央値 32.5）であるが，12 名中 9 名が平均と中央値のいずれも前掲の表 4 の「平均的」（21～30 点）の範囲内にある[6]。なお，評価者 12 名が 42 作文に対して行った総合評価点の基本統計量は，平均：26.0，中央値：25，最小値：2，最大値：50，標準偏差：11.96 であった。

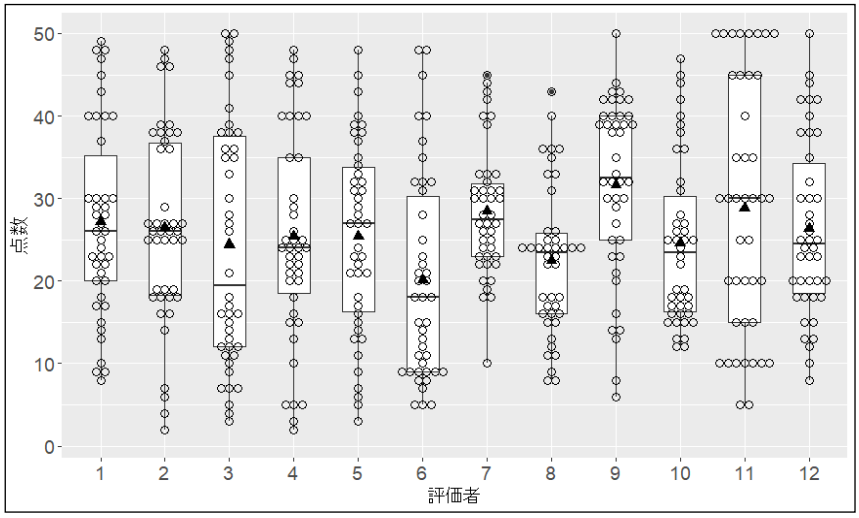

図 3　評価者別総合評価点の分布（12 名，42 作文）

[6] 評価者 3 は平均値 24.4，中央値 19.5 で，平均値は「平均的」（21～30 点）の範囲内であるが中央値はやや低い。

次に，評価者間の相関を見るためにスピアマンの順位相関係数を求めた[7]（表 5）。12 名の評価者から 2 名を抜き出すと組み合わせは全部で 66 組となる。66 組のうち，「強い相関あり（0.7<r≦1.0）」が 1 組，「中程度の相関あり（0.4<r≦0.7）」が 52 組，「弱い相関あり（0.2<r≦ 0.4）」が 13 組，「ほとんど相関なし（0<r≦ 0.2）」は 0 組で，66 組のうち 53 組（80.3%）に中程度以上の相関が見られた。なお，相関係数の解釈は山田・村井（2004）による。

表 5　総合評価点における評価者間の相関係数（N=12）

評価者	1	2	3	4	5	6	7	8	9	10	11	12
1	—											
2	0.63	—										
3	0.74	0.58	—									
4	0.43	0.53	0.62	—								
5	0.52	0.45	0.53	0.30	—							
6	0.39	0.44	0.58	0.56	0.24	—						
7	0.61	0.51	0.65	0.45	0.60	0.37	—					
8	0.54	0.56	0.59	0.53	0.25	0.58	0.40	—				
9	0.54	0.58	0.57	0.52	0.43	0.50	0.59	0.60	—			
10	0.34	0.35	0.51	0.42	0.25	0.48	0.21	0.57	0.33	—		
11	0.51	0.48	0.61	0.54	0.42	0.46	0.56	0.55	0.47	0.27	—	
12	0.50	0.56	0.65	0.66	0.42	0.66	0.59	0.64	0.57	0.36	0.45	—

図 3 および表 5 の結果からは，前掲の表 4 に示したように 42 本の作文の中で「中程度」を「21～30 点」の範囲としたことで，総合評価については評価者間の評価のずれをある程度小さくしたデータを収集することができたと考える。

RQ.2 の調査方法に関して，図 3 のプロットから 2 点指摘したい。まず，総合評価は 1 点刻みで評価を依頼したが，評価者 5 のように細かく点をつける人がいる一方で，評価者 11 のように 5 点間隔で評価する人もいたことがわかる。細かい点数を付けにくいと感じた可能性が考えられ，改善の必要性が示唆される。もう 1 点は，42 本全体に対して「平均的」な評価の範囲を決めても，特に 1～10 点（平均より非常に劣っている）の低い評価の作文が少なかったことである。この点については学習者の上達が早いことや，1 年生でもある程度整った作文を書ける学習者が多いことなどの理由も考えられる。しかし，評価者 7 と評価者 10 は，1～10 点の作文がそれぞれ 1 本（10 点）と 0 本である一方で，41～50 点（非常に優れている）の範囲の作文は複数本見られることから，低い評価をつけることに抵抗感を抱く評価者がいる可能性も考えられる。

6.2 項目別評価

次に，項目別評価（11 項目，5 段階評価）の値を一般化可能性理論（generalizability theory）を用いて検討する。一般化可能性理論は「古典的テスト理論で偶然誤差として包括的に扱っていた

[7] 相関係数を求める際にスピアマンの順位相関係数を用いたのは，データに正規性が仮定できなかったことによる。なお，分析には R（4.4.0 バージョン）を用いた。

ものを，問題項目に起因する誤差，評定者に起因する誤差等複数の誤差要因に分けて取り扱い，それらがテスト得点の分散の中でどの程度の大きさを占めるかを評価し，逆に，評定者を何名配置すれば所与の精度を持った測定が可能か，等について検討する」（近藤・小森（編）（2012）『研究社・日本語教育事典』: 338）分析手法である。スピーキングやライティングなどのパフォーマンス評価における評価項目の妥当性や評価の信頼性などを検討する際に多く用いられる（池田 1994，山森 2002，2004，山西 2005，水本 2008，平井 2018 など）。一般化可能性理論は，各要因および要因の組み合わせによる交互作用の分散成分を推定する G 研究（一般化可能性研究，Generalizability studies; G study）と，G 研究で算出された分散成分を用いて項目数や評価者数をシミュレーションする D 研究（決定研究，Decision studies; D study）の 2 段階から成る。本稿の「評価データの収集方法を検討する」という目的および RQ を明らかにするのに適した方法だと考え，分析方法に用いた。

なお，本稿の分析は池田（1994），山森（2002），水本（2008），平井（2018）を参考にし，誤差要因の分散を，分散分析の手法を適用した方法によって検討する。分析には，R（4.4.0 バージョン）の lme4 パッケージおよび GENOVA（Brennan 2016）を利用した[8]。なお，分析のデザインは 2 相完全クロス計画（池田 1994，山森 2002，GENOVA では，$p \times i \times r$ デザイン）である。

6.2.1 G 研究（一般化可能性研究）—変動要因の分散の検討

G 研究（一般化可能性研究）では「作文」「評価項目」「評価者」とこれらの交互作用を合わせた変動要因[9]が項目別評価全体に与える影響を見る。表 6 は「評価項目」（項目別評価 11 項目，5 段階評価）の評価データを用い，変動要因について，分散成分の推定値とその割合を求めた結果である。変動要因の「×」は交互作用を表す。例えば「作文×評価者」は作文と評価者の交互作用を指す。

表 6 の変動要因それぞれについて，項目別評価の評価値に占める分散成分の推定値において割合が高いのは「作文」「作文×評価者」「残差」でそれぞれ 36.1%，24.3%，31.6%であった。一方，「評価者」（0.5%），「項目」（2.6%），「作文×項目」（2.8%），「評価者×項目」（2.2%）はいずれも項目別評価の評価値に占める割合は低かった。

まず，「作文」（36.1%）は，分散成分の推定値の割合が最も高かった。これは作文によって評価が異なる，つまり高評価に偏る作文と低評価に偏る作文があることを表している。本調査では，初級学習者から日本語母語話者まで幅広いレベルの作文を評価の対象としているため，「作文」の割合が高いのは望ましいと言える。次に，「作文×評価者」（作文と評価者の交互作用）の割

[8] R の lme4 パッケージによる一般化可能性理論の実行方法は水本篤氏のホームページ（https://mizumot.com/handbook/?page_id=978/ 最終確認日:2024 年 8 月 28 日）参照。GENOVA の使い方は Brennan (2016) および平井 (2018) 参照。本稿では，R と GENOVA 両方を使用した。なお，R には一般化可能性理論を実行する gtheory パッケージもある（https://cran.r-project.org/web/packages/gtheory/gtheory.pdf 参照 最終確認日:2024 年 8 月 28 日）。

[9] 平井 (2018) によると，一般化可能性理論では，項目や評価者など系統的な誤差を生み出す要因を「相 (facet)」，測ることを意図している対象を「測定の対象（object of measurement)」と呼び，相と測定の対象と交互作用を合わせて，「変動要因（source of variability または source of variation)」と呼んでいる。本稿の印象評価の調査では，「評価項目」と「評価者」が「相」，「作文」が「測定の対象」に当たり，「評価項目」「評価者」「作文」とその交互作用を合わせたものが「変動要因」に当たる。

合も 24.3%と高かった。これは，項目別評価では同じ作文でも評価者によって評価の高低が異なるものがあることを示す。本稿では評価が一定の範囲内に収まる作文と，評価が分かれる作文があることを想定しているが，詳細は今後検討したい。「残差」については後述する。

表 6　項目別評価における分散成分の推定値とその割合

変動要因	自由度	分散成分	
		推定値	割合（%）
作文（*p*）	41	0.51	36.1%
評価者（*r*）	11	0.01	0.5%
項目（*i*）	10	0.04	2.6%
作文×評価者（*pr*）	451	0.34	24.3%
作文×項目（*pi*）	410	0.04	2.8%
評価者×項目（*ri*）	110	0.03	2.2%
残差（*pri,e*）	4510	0.44	31.6%
計		1.41	100.0%[10]

分散成分の推定値の割合が低かった「評価者」「項目」「作文×項目」「評価者×項目」については，各項目における中程度の作文を 5 段階評価の 3 としたため，評価の高低の偏りが小さかったことが影響していると考えられる。RQ.1「パイロット調査で安定した印象評価のデータが得られたか」について，評価者間の評価尺度のずれを軽減するという観点からみると，概ね望ましい結果と言える。しかし一方で，「評価者×項目」も低くなることから，今回の調査方法では，個人によって重視する項目が異なる場合，その実態は反映されにくいことが確認できる。例えば，評価者の中には，構成よりも正確さをより重視する人や，正確さよりも発想力をより重視する人もいると思われる。このような個人の中の評価観は現れにくい結果であるとも言える。

最後に「残差」について検討する。「残差」の割合は 31.6%と「作文」に次いで高かった。「残差」は，ここまで見てきた変動要因では説明できない要因である（Shavelson & Webb 1991，池田 1994，山西 2005，平井 2018）。要因が複雑であることから「残差」は検討されないこともある（山西 2005，平井 2018）が，「残差」の原因について，平井（2018:72）では「ある受験者がたまたま疲れてある項目だけ答えられなかった」可能性，日本語教師のライティング評価を分析した福田・石井（2016）では，評価者が評価基準を評価に十分に結び付けられなかった可能性，Shavelson & Webb（1991）では制限時間内に回答できなかった可能性などが挙げられている。これらの指摘を踏まえると，本パイロット調査において残差の割合が高いことからは，評価作業の方法や項目別評価の内容を検討する必要があることが示唆される。評価作業の方法については，筆者も同じ作

[10] 分散成分の割合を合計すると 100.1%になるが，これは各割合の小数点第 2 位を四捨五入したことによる。

業を行ったところ，三つの種類（体験文，説明文，意見文）が混ざった 42 作文をまとめて評価するのは作業負担が大きく，一気に作業するには慣れが必要だと思われた。

また，項目別評価の作業については，評価理由の記述から評価の観点を確認したところ，評価者の中に「結束性」の評価を「一貫性」の評価と混同している記述が見られた。表 7 は同一の作文（作文 B）に対する 1 人の評価者の記述である。

表 7　「結束性」と「一貫性」の評価理由（作文 B に対する評価者 5 の記述）

項目	評価	評価理由
結束性	3	テーマ（時計）について終始一貫した内容でブレていません。
一貫性	3	時計についての思い出を踏まえて，一貫した主張が展開されています。

前掲の表 3 に示したように，調査では「結束性」は「つながり」，「一貫性」は「まとまり」を見ている。評価者 5 の「結束性」の評価理由を見ると，「文と文は説明の部分では滑らかだが，結論部では不自然なつながりがある」のように，「結束性」の説明に合った記述も見られる。しかし一方で，評価者 5 の「結束性」の評価理由には，42 作文中 6 作文で「一貫」あるいは「まとまり」という語が使われており，そのほかにも「論点がブレずに記述されている」や「テーマに沿って記述されている」など，「一貫性」を判断していると思われる表現が記述されていた。評価者 5 のように「結束性」と「一貫性」を混同した記述はほかの評価者にも観察され，評価基準や説明に使う用語や表現の見直しが必要であることが確認された。

以上のように，G 研究の結果から RQ.2「安定したデータを得るために調査方法をどのように改善すればよいか」の示唆が得られた。

6.2.2　D 研究（決定研究）―評価者数のシミュレーション

RQ.3「1 作文当たり何人ぐらいの評価者に評価を依頼すれば，安定したデータが得られるか」を検討するために，D 研究（決定研究）を用いて分析を行った。D 研究では，G 研究で算出された分散成分の情報を用いて信頼性を求め，その値をもとに評価者数のシミュレーションを行う。一般化可能性理論での信頼性には 2 種類あり，一つは G 係数または一般化可能性係数（generalizability coefficient），もう一つはφ係数または信頼度指数（index of dependability）である。G 係数は受験者を順位づける場合などの相対的決定（relative decision）に用いられ，φ係数は，得点を絶対的決定（absolute decision）に利用するときに用いられる（平井 2018）。本稿では各作文が 5 段階の評価のうちどの段階に当てはまるかを見ているため，φ係数を用いて D 研究を行う。

まず，G 研究で得られた分散成分の推定値をもとに，G 研究を行った際（評価者数 12 名，11 項目）の ϕ 係数を求めたところ，φ=0.93 であった[11]。一般に信頼性の基準は 0.80 以上が望ましい（山森 2004，平井 2018）とされることから，本パイロット調査における評価方法は，一般化可能性理論の結果においては信頼性を十分に満たしたものであったと考えられる。

次に，評価者数のシミュレーションを行う。図 4 は G 研究で得られた分散成分の推定値を用い，評価者数と項目数が信頼性係数（本分析では ϕ 係数）に及ぼす影響をシミュレーションした結果である。本パイロット調査では信頼性係数が 0.80 を大きく上回る値であったため，評価者は 12 名以下を想定する。

図 4 は，評価者の人数ごとに項目数（x 軸）が変わると信頼性係数（y 軸）がどのように変化するかを示している。例えば，項目数が 11 の場合，評価者数が 4 名でも信頼性は 0.80 を満たす[12]。一方，人数ごとに結んだ線によって評価者数の傾向を見ると，評価者数が 7～8 名ぐらいまでは，人数ごとの線の間隔が比較的広く，それ以上になると線の間隔が狭くなる。そのため，7～8 名を超えたあたりから，評価者が 1 名増えたときの信頼性係数に及ぼす影響は小さくなることがわかる。例えば項目数 11 の信頼性係数を見ると，評価者数 6 名（0.875）と 7 名（0.890）の差は 0.015，7 名（0.890）と 8 名（0.901）の差は 0.011 であるが，8 名（0.901）と 9 名（0.910）の差は 0.009 で 0.01 を下回る。評価者数が多いほど信頼性は高まるが，評価者を 1 名増やすことによる調査実施にかかわる負担との兼ね合いを考慮して評価者数を決める必要があると思われる。

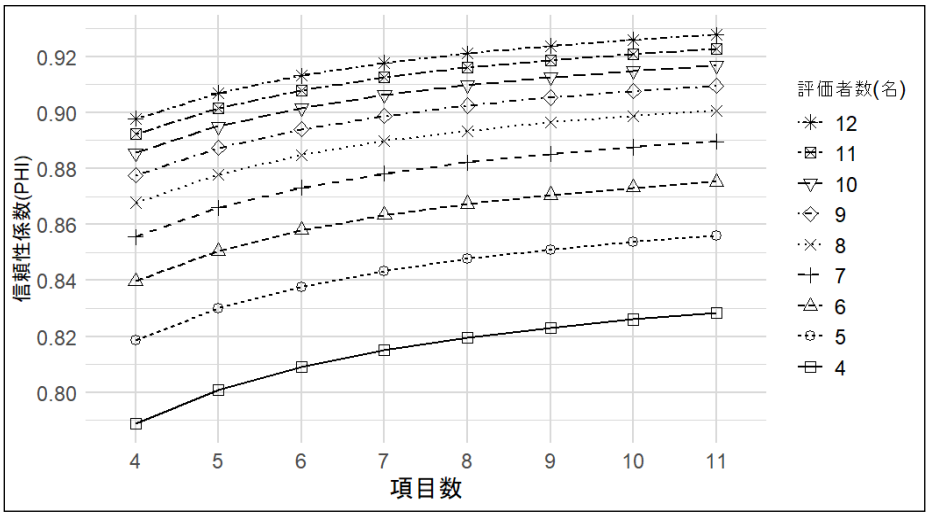

図 4　評価者数と項目数が信頼性係数（ϕ 係数）に及ぼす影響（シミュレーション）

なお，本稿で項目数の検討は RQ として設定していないが，図 4 では，評価者数のみならず項目数についてもシミュレーションが可能であるため，項目数についても述べておく。図 4 による

[11] ϕ 係数の計算は池田（1994: 36）の式を本研究のパイロット調査に当てはめて使用した。計算式は φ = 作文の分散/(作文の分散 +(項目の分散 +「作文×項目」の分散) / 項目数 +(評価者の分散 +「作文×評価者」の分散) / 評価者数 +(「項目×評価者」の分散 + 残差) / (項目数*評価者数))である。

[12] 評価項目数 11，評価者数 4 名の場合の信頼性係数は 0.83，図 4 には掲載されていないが，評価項目数 11，評価者数 3 名の場合の信頼性係数は 0.79 であった。

と，例えば，評価者数が 12 名の場合，5 項目以上あれば，信頼性は 0.90 以上となる。しかし，本プロジェクトでは，コーパスに幅広い観点からの印象評価の情報を付与するという目的があるため，項目数については，信頼性を参考にし，評価者の負担も考慮しながら，プロジェクトで別途検討することとしたい。

7.まとめと今後の課題

本稿では，日本語学習者縦断作文コーパス『W-CoLeJa』に付与する予定の印象評価データについて，安定したデータを収集するための方法を検討した。データはクラウドソーシングサービスを通じて収集したパイロット調査の結果である。分析には総合評価と項目別評価のデータを用い，総合評価の分析には，相関分析，項目別評価の分析には，一般化可能性理論の G 研究および D 研究を利用した。以下，本稿の結果を RQ に沿ってまとめる。

RQ.1「パイロット調査で安定した印象評価のデータが得られたか」については，中程度の作文についての点数の目安を設定することによって，総合評価と項目別評価のいずれにおいても，評価者間の評価尺度のずれをある程度小さくしたデータが収集できたと思われる。

RQ.2「安定したデータを得るために調査方法をどのように改善すればよいか」については，次の 3 点が課題として挙がった。1 点目は，総合評価について 1〜50 点の 1 点刻みの評価は細かすぎるため，10 段階評価などの粗い評価にしたほうがよいということ，2 点目は 42 作文まとめて評価するのは作業負担が重いため，一度に評価を行う作文数を減らしたほうがよいことである。そして 3 点目は「結束性」のような専門的な用語は，一般の日本語母語話者には理解しにくい場合があるため，説明に用いる語はより平易な言葉に置き換える必要があるということである。これらの改善点はいずれも基本的なことではあるが，調査者が一度に多くのデータを取ろうとすると，作業者の作業にかかる負担や作業の困難度は見過ごされやすいことを表していると思われる。

RQ.3「1 作文当たり何人ぐらいの評価者に評価を依頼すれば，安定したデータが得られるか」については，一般化可能性理論の D 研究の結果によると，項目数が 11 項目の場合，評価者数 4 名以上で信頼性は 0.80 を満たす。また，評価者が増えることによって，信頼性係数は高くなるが，7〜8 名を超えたあたりから，評価者が 1 名増えたときの信頼性係数に及ぼす影響は相対的に小さくなることも確認された。これらのことから，評価者数はパイロット調査より減らしてもある程度の信頼性が確保されることがわかった。

本稿の分析方法の課題として挙げられるのは，調査対象の作文には「体験文」「説明文」「意見文」の三つの種類の作文を合わせた全体を一つの作文データとして扱っていることである。項目別評価の中には，作文の種類によって評価のしやすさが異なるものが含まれる可能性があるため，評価の理由の記述を利用しながら確認する必要がある。また，今後，本格的に調査を行うにあたって検討が必要なのは，調査対象の作文の選び方である。「中程度の作文」は，調査対象の作文の選定方法に依存する。そのため，まとまった量の作文を複数回に分けて調査を実施する際には，すべての回に共通する作文を入れて，各回の評価作業に用いる作文のレベルを確認するなど，作業方法にさらに工夫が必要だと考える。

本稿の分析結果を生かしてより安定したデータの収集を目指し，将来的には収集したデータを作文の自動添削システムなどの評価ツールの構築に活用したい。

参照文献

Shavelson, Richard J. and Noreen M. Webb (1991) *Generalizability theory: A primer*. CA, US: Sage Publications.

池田央 (1994)『現代テスト理論』東京：朝倉書店.

石黒圭 (2017)「文章とは何か 日本語の表現面から見たよい文章」李（編） (2017), 14–37.

伊集院郁子 (2017)「作文と評価 日本語教育的観点から見たよい文章」李（編） (2017), 38–57.

伊集院郁子・李在鎬・小森和子・野口裕之 (2020)「評価コメントに見られる意見文評価の様相―共起ネットワーク及びコレスポンデンス分析に基づく考察」『第二言語としての日本語の習得研究』23: 26–43.

宇佐美洋 (2014)『「非母語話者の日本語」は、どのように評価されているか』東京：ココ出版.

近藤安月子・小森和子（編） (2012)『研究社・日本語教育事典』東京：研究社（「一般化可能性理論」執筆者：野口裕之・大隅敦子）.

田中真理 (2016)「パフォーマンス評価はなぜばらつくのか？ アカデミック・ライティングの評価における評価者の「型」」宇佐美洋（編）『「評価」を持って街に出よう』34–53. 東京:くろしお出版.

田中真理 (2022)「ライティング評価の限界といいとこ取り」鎌田修・由井紀久子・池田隆介（編）『日本語プロフィシエンシー研究の広がり』225–237. 東京：ひつじ書房.

田中真理・坪根由香里・初鹿野阿れ (1998)「第二言語としての日本語における作文評価基準―日本語教師と一般日本人の比較―」『日本語教育』96: 1–12.

田中真理・長阪朱美 (2006)「第 2 言語としての日本語ライティング評価基準とその作成過程」国立国語研究所（編）『世界の言語テスト』253–276. 東京：くろしお出版.

田中真理・長阪朱美 (2009)「ライティング評価の一致はなぜ難しいか―人間の介在するアセスメント―」『社会言語科学』12(1): 108–121.

平井明代 (2018)『教育・心理・言語系研究のためのデータ分析』東京：東京図書.

福田純也・石井雄隆 (2016)「中国語を第一言語とする日本語学習者の作文に対する日本語教師の評価：一般化可能性理論を用いた検討」『日本教科教育学会誌』39(2): 81–89.

水本篤 (2008)「自由英作文における評定者評価の種類と信頼性」統計数理研究所共同研究リポート 215『学習者コーパスの解析にもとづく客観的作文評価指標の検討』43–49. 東京：統計数理研究所.

山田剛史・村井潤一郎 (2004)『よくわかる心理統計』京都：ミネルヴァ書房.

山西博之 (2005)「一般化可能性理論を用いた高校生の自由英作文評価の検討」『JALT Journal』27(2): 169–185.

山森光陽 (2002) 「一般化可能性理論を用いた観点別評価の方法論の検討」『STEP Bulletin』14: 62–70.
山森光陽 (2004) 「英会話テストの信頼性の検討―一般化可能性理論―」 前田啓朗・山森光陽（編著）『英語教師のための教育データ分析入門―授業が変わるテスト・評価・研究』82–89. 東京：大修館書店.
李在鎬（編） (2017) 『文章を科学する』東京：ひつじ書房.

分析ツール

Brennan, Robert L. (2016) GENOVA (University of Iowa, Center for Advanced Studies in Measurement and Assessment (CASMA)). https://education.uiowa.edu/casma/computer-programs (最終確認日:2024 年 8 月 29 日).

Methods for Collecting Stable Data on Impression Evaluations of Japanese Learners' Writings

HONDA Yumiko[a] II Nahoko[b]

[a] Research Department, NINJAL

[b] University of the Ryukyus/ Project Collaborator, NINJAL

Abstract

This study explores stable data collection methods for including impression evaluation data in the longitudinal corpus of Japanese learners' writings, "W-CoLeJa". In the pilot survey, 12 native Japanese speakers evaluated 42 writings, and the results were analyzed using correlation analysis and generalizability theory (G theory), yielding three findings. First, by using a medium level of writing as a benchmark, the deviation in the evaluation scale among raters could be reduced, enabling the collection of relatively stable data. Second, improvements to the collection method include changing the overall rating from a fine-grained 50-point scale to a coarser one, such as a 10-point scale, reducing the number of writings evaluated all at one time, and replacing technical terms such as "cohesion" in the explanations with simpler words. Third, regarding the number of raters, the results of a decision study (D study) based on G theory suggest that even if the number of raters is reduced in future surveys, maintaining a certain level of reliability with the current 11 evaluation items is possible.

Keywords: W-CoLeJa, writing, Japanese native speakers, crowdsourcing, generalizability theory