印象評価情報の付与：クラウドソーシングの活用

第13章｜印象評価情報の付与
―クラウドソーシングの活用―

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」では、『日本語学習者縦断作文コーパス（W-CoLeJa）』に収録する学習者の作文に印象評価情報を付与するため、クラウドソーシングを利用して一般の日本語母語話者を対象に作文の印象評価調査を行っている。本章では印象評価の調査を設計する際の試行錯誤の過程と本調査で収集されたデータを紹介し、今後の印象評価データ活用の可能性について述べる。

KEYWORD
｜学習者作文｜日本語母語話者｜評価項目｜評価尺度｜評価者数｜

□1　印象評価情報付与の意義

よく書けている作文とはどのような作文だろうか。一般に作文の評価と言うと、学期中や学期末に提出された作文を教師が採点するような場面が想起されるかもしれない。授業の一部として行われる評価では、文法や語彙、表記の誤用をチェックするといった正確さに対する評価が中心となる傾向がある。一方で、教育以外の場面では、就職活動のエントリーシートや知人に近況を伝えるメールなど、日本語教師以外の日本語母語話者が学習者の作文の読み手となることは少なくない。そこでは、「内容に共感した」や「おもしろい文章だ」といった作文全体から受ける印象も含めた評価がなされると思われる。

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研究は、評価者が教師や教員に限られており、分析の材料となるデータの蓄積も十分とは言えない。このため、一般の日本語母語話者から学習者の作文に対する印象を含めた評価データを付与することはライティング教育に資する研究として意義があると考える。そこで、私たち「日本語学習者の作文の縦断コーパス研究」プロジェクト（以下「本プロジェクト」）では、作文から受ける印象を評価することを「印象評価」と呼び、クラウドソーシングを活用し、一般の日本語母語話者から印象評価データを収集することにした。以降では、収集した印象評価

データの概要とその収集過程を紹介する。本章の執筆者である本多は調査の設計・実施とデータ分析を、井伊は国立国語研究所にプロジェクト非常勤研究員として在職した当時、調査の設計・実施を担当し、石黒はプロジェクトリーダーとして調査の企画・設計に中心的に携わっている。また、調査実施の際、鈴木英子氏（元技術補佐員）、江澤実紀氏（プロジェクト非常勤研究員）にご協力いただいた。なお、本章は本多（2024）、本多・井伊（2025）、井伊（2025）をもとにしている。

□2　先行研究と印象評価情報付与における課題

2.1　先行研究

本節では、日本語学習者の作文に対する評価が「誰／何による評価か」に注目し、人間による評価とシステムによる評価について述べる。まず、人間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（名古屋外国語大学）を中心とする「GoodWriting」の一連の研究（田中・長阪 2006, 2009; 田中ほか 2009; 田中・坪根 2011; 田中 2016, 2022 など）や伊集院（2017）、伊集院ほか（2020）などがある。これらの研究では、作文評価のための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; 田中ほか 2009）、インタビュー・評価コメントなどを用いた教師間の評価観点の相違の分析（田中・坪根 2011; 田中 2016; 伊集院 2017; 伊集院ほか 2020）などが行われ、教師によるライティング評価を支援する取り組みが進められている。

一方で、一般の日本語母語話者を評価者とした研究は田中ほか（1998）や宇佐美（2014）などに限られる。田中ほか（1998）では、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調査が行われ、両者の重視する点には相違があり、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、日本語母語話者による評価の多様性を「評価プロセスモデル」を作成することで可視化している。

ライティング評価のようなパフォーマンスの評価では、評価者の主観が評価に影響する。そのため、評価者個人における評価の揺れ（McNamara 2000）や、複数の評価者による評価の不一致（田中 2016, 2022; 宇佐美 2014）が課題として挙げられる。評価の揺れや評価者間のずれを軽減するためには評価基準の設定や評価者のトレーニングが必要だとされる（McNamara 2000; 近藤 2012）。

次に、システムによる評価では、近年「GoodWriting Rater」や「jWriter」などの自動評価システムが開発、公開されている。これらは、学習者コーパスのデータを活用したもので、「GoodWriting Rater」には作文データの定量的な数値と日本語教師による作文への評価スコアをもとに作られた計算モデルが用いられ（田中 2022）、「jWriter」には作文データの定量的な数値と学習者の日本語能力を測るテスト結果とを組み合わせて作られた計算モデルが用いられている（李ほか 2019）。自動評価システムは前述のような評価の揺れの問題を軽減し、ライティング教育を支援するツールとして、今後の進展が期待される。

2.2　印象評価情報の付与における課題

前述のように、人が作文に対して評価を行う場合、各評価者内の評価の揺れや、評価者間の評価が一致しにくいことが課題となる。本プロジェクトにおいて測ろうとする、日本語母語話者が作文を読んだ後の印象も、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を通じて1本の作文に対し複数名から収集した評価値の平均を求めることによって、評価者間の評価の差をならすことにした。また、複数の評価者間の評価尺度のずれを小さくするためにはどのような調査方法・調査項目を設定すればいいかを検討することを目的に、パイロット調査を実施した。次節ではこのパイロット調査について、どのような試行錯誤を経て調査設計を固めたのかを記述する。

□3　印象評価調査の設計に向けたパイロット調査

3.1　パイロット調査の概要

本プロジェクトでは本調査に先立ってパイロット調査を3回実施した。本節では、パイロット調査での検討課題とその検討結果について述べる。

まず、パイロット調査の概要を次ページの表1に示す。3回のパイロット調査を通して五つの検討課題「①評価者間の評価尺度のずれを小さくできるか」「②何名の評価者に依頼をするか」「③項目別評価の評価項目に問題がないか」「④総合評価の数値をどのように設定するか」「⑤評価理由をどのように記述してもらうか」が生じた。それぞれ節を分けて取り上げる。

表1　パイロット調査の概要

回数	年月	作文1本あたり評価者数	作文数	内訳			
				母語※1	学年※2	テーマ※3	本数
1	2023.01	3名	9	中（簡）・韓・越	1~3	A1~A3	※4
2	2023.03	16名	15	中（簡）・越・日	1~2	A1~A3	各1
3	2023.11	13名	42	中（簡）・越・日	1~3	A1~A3	各2

※1　中（簡）：簡体字圏、韓：韓国語圏、越：ベトナム語圏、日：日本語母語話者
※2　日本語母語話者は1学年分のみ
※3　テーマの詳細は第1章参照
※4　学年とテーマがばらけるように中（簡）4本、韓2本、越3本を選定した

3.2　検討課題①評価者間の評価尺度のずれを小さくできるか

クラウドソーシングで募集した一般の日本語母語話者が調査協力者の場合も、評価者間の評価尺度のずれを小さくすることはできるのだろうか。

本プロジェクトでは前述の通り、クラウドソーシングを通じて作文の印象評価データを収集している。クラウドソーシングとは、オンライン上で仕事をしてほしいクライアントと仕事をしたいワーカーを繋いでくれるサービスのことであり、本プロジェクトではクラウドソーシングサービスを提供している株式会社クラウドワークス（https://crowdworks.jp/）を利用した。そもそもクラウドソーシングを用いたのは、日本語教師ではない一般の日本語母語話者を中心に印象評価データを収集することを目指しており（本章1節参照）、一般の日本語母語話者を募るのにクラウドソーシングが適していると考えたためである。

しかし、クラウドソーシングを利用して印象評価データを収集する場合に心配されたのは、属性も経験も仕事に対する取り組み方もさまざまなワーカーが集まる場で、同じような評価尺度で評価することが可能かという点であった。

そこで、パイロット調査3回目から、評価者間における評価尺度のずれを小さくするために「中程度」という基準を設けた。評価尺度を示したものが表2である。まず評価者に配布した調査対象の作文のセットの中で中程度の評価だと思う作文を「平均的」（項目別評価であれば5段階のうちの3）として点数をつけてもらい、その中程度の作文を基準とした相対評価で評価値をつけてもらうという評価方法である。この基準がなかったパイロット調査1回目と2回目では評価者によって評価の厳しさに差が観察され、低い評価や高い評価に偏る評価者がいたのに対し、パイロット調査3回目で中程度という基準を設けた結果、評価者12名中9名の総合評価が平均値・中央値いずれも「平均的」の範囲内に収まった。また、12名の評価者間の相関をみるためにスピアマンの順位相関係数を求めた

ところ、66組（12名の評価者から2名ずつ抜き出した組み合わせ数）のうち80.3%にあたる53組に中程度以上の相関が観察された（詳細は本多・井伊（2025）参照）。以上から、中程度という基準によって評価者間の厳しさの違いによる評価尺度のずれをおさえることができたと考え、この基準を本調査でも使用している。

表2　本調査の評価尺度

評価の目安	総合評価（10段階）	項目別評価（5段階）
平均より非常に優れている	9〜10	5
平均よりやや優れている	7〜8	4
平均的（中程度の作文）	5〜6	3
平均よりやや劣っている	3〜4	2
平均より非常に劣っている	1〜2	1

なお、中程度という基準は、評価対象としてどのような作文を選定するかに依存する。本プロジェクトでは評価対象の作文数が多く、作文を複数のセットに分けて評価者に割り振る必要があるため、各セットに1年生から4年生までの作文を、母語の偏りがないようにバランスを考慮して配分した。また、すべてのセットに共通する作文（以下、「共通作文」）を入れ、各セットの評価対象の作文のレベルに大きな差がないかを調査実施後に確認することにした。本調査における共通作文の評価の結果については、4.3で検証を行う。

3.3　検討課題②何名の評価者に依頼をするか

検討課題の二つ目が評価者の人数である。1作文あたりの評価者数が多いほど個人差の影響をおさえることができ、データの信頼性が高まる一方、調査実施にかかるコストも増加する。そこで何名の評価者に依頼すれば信頼性の高いデータが得られるかをパイロット調査3回目のデータを用いて検討した。

本多・井伊（2025）において一般化可能性理論のD研究（決定研究）による統計的なシミュレーションを行った結果、評価者が7〜8名を超えると評価者が1名増えたときの信頼性係数に及ぼす影響が小さくなることがわかった（詳細は本多・井伊（2025）参照）。これを踏まえ、本調査では、1作文あたり評価者7名分の印象評価データを収集することを目安としている。

3.4　検討課題③項目別評価の評価項目に問題がないか

検討課題の三つ目からは評価項目に関する検討課題を扱う。まず、本調査で

採用した評価項目を表3に示す。評価項目には総合評価・全体の理解・項目別評価の3種があり、項目別評価は11項目に下位分類される。

表3　本調査の評価項目

項目		評価内容
総合評価（10段階評価）		よく書けているかどうかという観点から作文を評価する。主観的な評価でよい。中程度の作文を「平均的（5〜6）」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）		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. 一貫性	文章の筋が通っていてまとまりがある。
	8. 文のつながり	前後の文のつながりがなめらかである。
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。

上記の表3の評価項目にまとまるまでに、「検討課題③項目別評価の評価項目に問題がないか」について検討を行った。修正が生じたのは、「結束性」という名称だった項目を「文のつながり」に変更した点である。パイロット調査1回目〜3回目で項目別評価の「構成」の一つとして「結束性（文と文のつながりがなめらかで自然である。）」という項目を設けたところ、「結束性」の評価の理由として「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。」（評価者5、評価3）のように同じ「構成」の項目である「一貫性」と混同している様子が観察された。これは一般の日本語母語話者にはなじみの薄い専門用語であることに問題があったと考えられる。そこで、本調査からは「結束性」を「文のつながり」という身近な用語に変更している。

3.5　検討課題④総合評価の数値をどのように設定するか

評価項目に関連するもう一つの検討課題が、「総合評価の数値をどのように

設定するか」である。プロジェクト内では二つの意見に分かれていた。一つは総合評価を1〜10の10段階とする意見、もう一つは作文間の評価の差が詳細に現れるように1〜50点（1点刻み）とする意見である。

そこで、パイロット調査3回目で1〜50点評価を試行した。その結果、1点刻みで点数をつけた評価者と、評価者11のように5点刻みで点数をつけた評価者に分かれ（図1）、相対的に評価するには1〜50点評価は点数が細かすぎることがわかった（詳細は本多・井伊 2025を参照）。そのため、本調査の総合評価は10段階評価としている（前掲の表2参照）。

なお、図1のように5点刻みの評価者がいたのにはもう一つの理由が考えられる。それは、パイロット調査3回目では42本の作文内での相対評価を依頼したため、評価する作文数が多すぎたということである。そこで、本調査では作文14本を1セットとし、評価する作文数を減らして実施している。

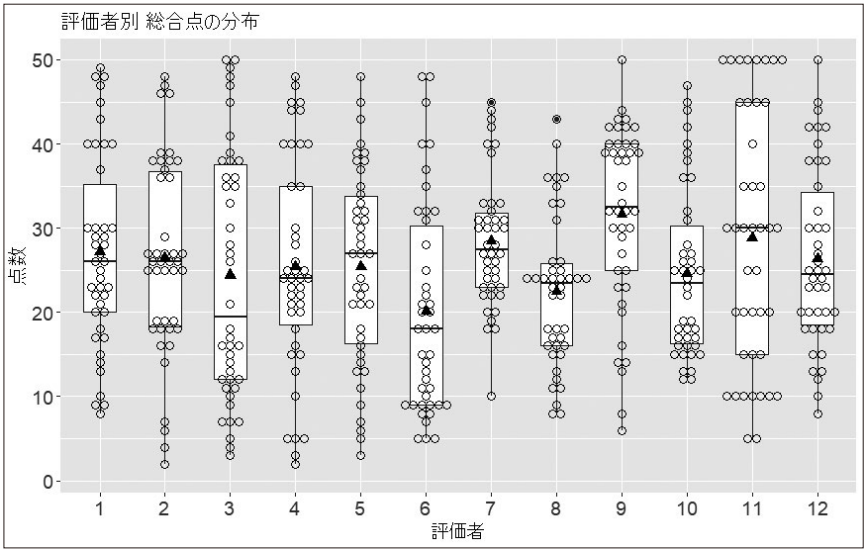

図1　パイロット調査3回目の評価者別総合評価点の分布（本多・井伊 2025）

3.6　検討課題⑤評価理由をどのように記述してもらうか

作文の印象評価の分析では、評価の数値だけでなく評価の理由も重要になる。そこで、「評価理由をどのように記述してもらうか」についても検討を行った。

はじめは適切に指示が伝わらず、評価の理由が明確に書かれない問題が生じ

た。たとえば、「各項目の評価の理由を書いてください」と指示したところ、「説得力5」の評価理由が「説得力があると思ったから」という具合である。

そこで、本調査では、「〈理由〉それぞれの項目の評価の理由を具体的に書いてください。」と指示したうえで、書き方の「よい例」と「よくない例」を提示した。提示した「よい例」は「……と書かれている点が……だと感じた。」「『……』のように……ている点が……だと感じた。」、「よくない例」は「『2. 説得力』の項目について『説得力があまり感じられませんでした』」である。

3.7　誤解析につながる表記・活用の誤りの事前修正

前項まではパイロット調査を通して検討・修正してきた事柄を取り上げてきたが、本節の最後にパイロット調査開始時から本調査まで一貫して実施してきた、誤解析につながる表記・活用の誤りの事前修正について取り上げる。

本プロジェクトでは、作文を形態素解析した際に誤解析になる恐れのある表記・活用の誤りが作文に含まれていないかを複数人で確認し、もし発見された場合は事前に修正したうえで印象評価調査に使用している。このような方針をとった理由の一つは、AI技術の進歩がめざましい現状を鑑みると、近い将来、作文に含まれる表記や活用などの明らかな誤用はAI技術などの機械処理で自動的に修正され、人の印象評価を受けるのは修正後の作文になることが予想されるためである。もう一つの理由は、本印象評価調査では、日本語学習者が書いた作文に含まれる日本語の誤用に対する評価だけでなく、作文を読んだ後の印象（おもしろいと思えるか、共感できるかなど）も含めた総合的な評価のデータを収集することに重きを置いているためである。日本語学習者の作文を読み慣れていない一般の日本語母語話者が日本語の誤用に引きずられすぎないよう、表記・活用の誤りのみ事前修正を行った。

なお、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箇所であった（パイロット調査3回目のデータに基づく。日本語母語話者の作文を除く）。この事前修正は本調査でも引き続き実施している。

以上がパイロット調査を通して検討してきた事項である。次節からは、パイロット調査の検討を踏まえて行った本調査について述べる。

□4　本調査で収集した印象評価データの概要

本調査では、2025年4月現在、240本の作文に対する印象評価データを収集した。以下では、これまでに収集したデータの概要を示し評価結果を概観する。

なお、W-CoLeJaで公開される印象評価データは、本調査で収集した印象評価データと評価対象の作文が一部異なる。本章の分析に用いた、以下の表4に示す作文の印象評価データは国立国語研究所のリポジトリ（https://doi.org/10.15084/0002000581)で公開する。

4.1　評価対象の作文

評価対象の作文は、W-CoLeJaの中から中国語（簡体字圏)、中国語（繁体字圏)、ベトナム語圏の大学で学ぶ日本語学習者各10名（計30名）の4年間の作文が選ばれた（表4)。第1回（2024年2〜3月）の調査では30名の1テーマ（A1「うれしかったプレゼント」）の4年分の作文を対象に調査を行い、第2回（2025年2〜3月）では第1回の調査対象となった学習者の中から各5名を選んで、2テーマ（A2「好きな有名人」とA3「写真と動画」）について4年分の作文を対象に調査を行った（テーマについては第1章参照)。調査対象の作文は各回の作文執筆で一定の期間内に提出された作文であること、特定の大学に偏らないことなど、複数の選定基準が設けられた。また、実際の印象評価調査では、第1回、第2回ともに日本語母語話者の大学1年生が同じテーマで執筆した作文を比較対象として10本ずつ加え、130本の作文が評価対象となった。

表4　評価対象の作文

母語	第1回				第2回				作文数計
	人数	テーマ	学年	作文数	人数	テーマ	学年	作文数	
中国語（簡）	10	A1	1〜4	40	5	A2,A3	1〜4	40	80
中国語（繁）	10	A1	1〜4	40	5	A2,A3	1〜4	40	80
ベトナム語	10	A1	1〜4	40	5	A2,A3	1〜4	40	80
日本語	10	A1	1	10	5	A2,A3	1	10	20
計	40	—	—	130	20	—	—	130	260

・（簡）：簡体字圏、（繁）：繁体字圏

4.2　評価者数およびデータの収集方法

第1回、第2回ともに評価対象の作文130本（学習者作文120本、日本語母語話者作文10本）に対し、作文1本あたり7名の評価者による評価データを収集する目的で調査が行われた。しかし、第1回の調査で評価者のうち1名が途中で評価作業を辞退したため、第1回の調査では、作文1本あたりの評価者数が6名または7名となった。評価者は1セット14本の作文を読み、評価を行った。各セットはそれ

ぞれ似たようなレベルの作文から成るように、まず中国語（簡体字圏）、中国語（繁体字圏）、ベトナム語圏の各学年から1本ずつランダムで12本が選択された。そこに、日本語母語話者による作文1本と、共通作文が1本加えられた。共通作文は、パイロット調査で中程度の評価を受けた作文の中から各テーマ1本ずつ選ばれたもので、調査後に各セットのレベルを確認する目的で加えられた。

評価者は第1回と第2回の実施時にそれぞれ募集された。前述のように途中で辞退した1名を除く延べ27名のうち、9名が第1回と第2回の両方で評価を行った。表5に自己申告による年齢、性別、職業内訳を示す。なお、日本語教育に携わった経験があると回答した人は第1回、第2回ともに1名であった。本プロジェクトでは一般の日本語母語話者の中にも日本語非母語話者との接触経験を有する人は少なくないと考え、職業が日本語教師でなければ日本語教育に携わった経験の有無は問わないことにした。

表5　評価者の属性

属性	第1回（13名）	第2回（14名）
年齢	20代1名、30代3名、40代6名、50代2名、60代1名	30代4名、40代9名、50代1名
性別	男性6名、女性7名	男性6名、女性8名
職業	会社員4名、専業主婦2名、フリーランス2名、ITエンジニア、アルバイト、パート社員、高校の非常勤講師、法律事務所事務員各1名	会社員4名、フリーランス/自営業4名、専業主婦3名、アルバイト、パート、教員各1名

4.3　共通作文の評価

評価値を集計する前に、すべてのセットに1本ずつ含まれる共通作文の総合評価の値を確認する（表6）。なお、すべてのセットにおいて、14本の作文の総合評価は最小値が1で最大値は10であった。また、本章での計算にはR（ver.4.4.2）を用いた。

A1～A3について、多少のばらつきは見られるが、クラスカル・ウォリス検定の結果、各セットにおける共通作文の評価値に有意差は認められなかった。次ページの表6の結果から、テーマごとの調査で用いられた各セットの作文はそれぞれ似たようなレベルの作文から成るものと判断し、評価値を用いることとしたい。

表6　各セットにおける共通作文の総合評価（10段階評価）

テーマ	セット数	平均値の範囲	中央値の範囲	標準偏差の範囲	検定結果
A1	10	4.67〜5.43	5〜6	1.11〜1.94	H (9) = 2.60, p=.98,ns
A2	5	4.85〜5.71	5	0.53〜1.25	H (4) = 2.80, p=.60,ns
A3	5	4.57〜6.29	5〜7	1.85〜2.39	H (4) =3.24, p=.52,ns

※検定結果はクラスカル・ウォリス検定による。

4.4　収集したデータ

W-CoLeJaは日本語学習者から同じテーマの作文データを4年間収集している点に特徴がある。そこで、A1「うれしかったプレゼント」に対する4年間の評価を取り上げる。まず、総合評価について4年間の推移を見たあとで項目別評価との関係を概観する。次に、評価理由について「説得力」を取り上げる。

4.4.1　総合評価の推移

図2では各学習者の総合評価の平均を母語別に箱ひげ図にまとめた。

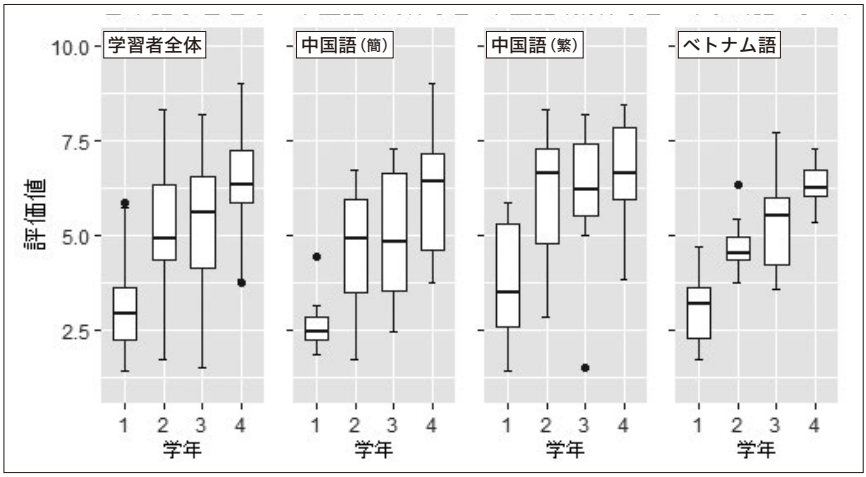

図2　総合評価（平均）の推移

日本語学習者全体では、評価値の伸びは1年生と2・3年生、4年生の3つに分かれるように見える。まず、1年生は全体的に評価値が低いが1年生から2年生で大きく伸びている。2・3年生では、中央値に上昇は見られるものの各学年における評価値の上下の範囲に大きな変化は見られない。そして、4年生で全体

的に評価が上がっている。これを母語別に見ると、学年ごとの上下の幅が比較的小さいのはベトナム語を母語とする学習者である。また、中国語（繁体字圏）では1年生から2年生で評価が上がり、その後の伸びは小さい。なお、調査では日本母語話者のデータも収集したが1学年のみのデータであるため、図2からは除いている。

4.4.2　個人における総合評価の推移

全体的な傾向として4年間で評価が上昇していることはわかったが、これを個人別に確認してみたい。図3は図2で示した4年間の総合評価の平均の推移を学生ごとに折れ線グラフで示したものである。1年生と4年生を比較すると全員評価は上がったが、グラフ中の実線のように、評価が上昇し続ける人だけではなく、グラフ中の点線のように、4年の間に評価が下がる人も見られる。このグラフではベトナム語母語話者よりも中国語（簡体字圏）母語話者のほうに2年生から4年生の間で評価が下がる人が目立つ。

評価が下がった理由については複数の可能性が考えられる。新しく学んだ表現を使ったものの、意味がうまく伝わらなかったことで評価が下がる場合がある。また、学習者が作文の執筆に割いた時間や日本語学習に対するモチベーションの変化など文章からは見えない要因が影響している可能性も考えられる。

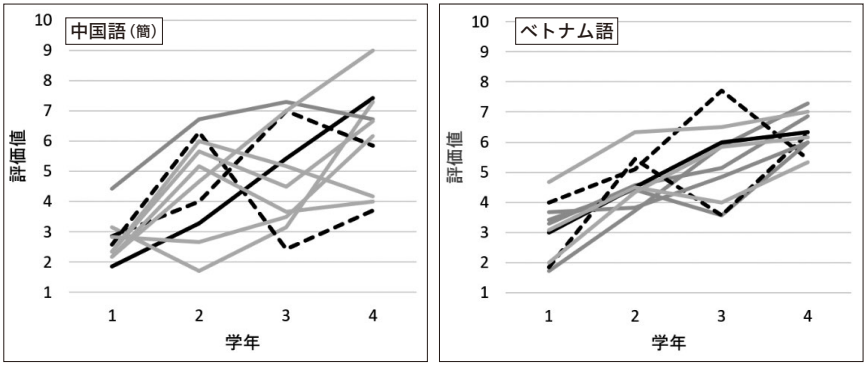

図3　総合評価（平均）の個人別推移

4.4.3　総合評価と各評価項目との相関

ここでは、総合評価と項目別評価の関係性を相関分析の結果から把握しておく。表7には、A1「うれしかったプレゼント」（130本）の作文における総合評価

と各評価項目の値を用いてスピアマンの相関係数を求めた結果を示す。すべての項目に総合評価との「強い相関（0.7＜r≦ 1.0）」が見られたが、その中で最も高かったのは説得力（0.85）で、最も低かったのは正確さ（0.70）であった。評価者がどのような要素に注目しているかを検討する必要があるが、表現の正しさよりも文章から受ける「なるほど」という印象のほうが総合評価との関係が強いことがわかる。

表7　総合評価と各評価項目との相関係数（N=845）

	項目	相関係数
内容	発想力	0.78
	説得力	0.85
	情報面での配慮	0.84
	心理面での配慮	0.76
	共感力	0.82

	項目	相関係数
構成	構成	0.81
	一貫性	0.77
	文のつながり	0.76
表現	正確さ	0.70
	ふさわしさ	0.75
	豊かさ	0.80

4.4.4　評価理由

項目別評価には評価の理由も記載されている。たとえば、下記の作文例1はベトナム語圏の2年次の学習者が書いたA1「うれしかったプレゼント」の作文である（下線は筆者による）。この作文について「説得力」の評価結果を見ると、評価が1の評価者もいれば4の評価者もいる。同じ作文を読んでも評価がばらつくのはなぜだろうか。評価がばらつく原因を評価の理由から探ってみたい。

（作文例 1）{前略} 大学に入学した時、姉にバイクをもらいました。それはもらってうれしかったプレゼントです。最初はバイクに乗った時、ちょっと心配しましたが、今はうまく乗れるようになりました。このバイクのおかげで、家から学校まで通う時間も少しかかるし、バスを待って、疲れたのもないし、それで便利で、役に立つと思います。そして、{中略。筆者注：色々なところに行けることが具体的に書かれている} ほかに、バイクは車に比べて、安いし、使い方が簡単だし、それで、学生たちに会います。でも、バイクのせいで、最近はあまり歩かずにいるので、体が太くなりました。そして、前ほど早く起きません。ですから、できるだけ、一所懸命歩きます。（ID：VVA025-A1-02-04）

評価の理由を確認すると、下記の表8のような記載が見られた。ここでは紙幅の都合上、2名分のみ掲載する。

表8 「説得力」の評価理由の例（5段階評価）

項目	評価	評価理由
説得力	4	バイクをもらってどうして嬉しかったのか、具体的な理由が書かれていて納得できた。（評価者7）
	1	バイクが必要な理由が明確に書かれておらず、説得力に欠けた。（評価者6）

表8を見ると、「説得力」を高く評価した評価者も低く評価した評価者も「理由」というキーワードが含まれている。しかし、「どうして嬉しかったのか」に着目する評価者7と「バイクが必要な理由」に着目する評価者6とで理由の捉え方が異なっており、評価値も異なる結果になっている。「バイクが必要な理由」に着目した評価者6は、作文の最後の下線部に、バイクには乗らずできるだけ徒歩で移動するようになった経緯が書かれているため、「説得力」が低いと判断したものと思われる。「説得力」では、理由の捉え方が異なることで評価結果がばらつく場合があることがわかる。

□5 印象評価情報の活用の可能性

最後に、印象評価情報を作文に付与することによって今後どのような活用の可能性があるかについて述べる。

まず、教育現場の視点から作文教育への応用が考えられる。多くの人が良い印象を持つ作文や評価がばらつく作文の特徴を知ることができれば、教育現場の日本語教師や日本語学習者にとって有益なものになるだろう。たとえば、調査結果に基づいてルーブリックを作成し、学習者の自己評価の指標にも活用できる可能性がある。さらに、本プロジェクトでは一般の日本語母語話者の評価を収集しているため、教育現場以外のSNSの文章などにおいてどのような文章が良い印象を持たれやすいかを知るための指標にもなると考えられる。

また、研究の視点から、作文そのものの言語学的な研究や評価の研究にも活用できると考えられる。とくに評価の研究としては、評価者と評価値の関係や、評価者の属性との関係、作文の自動評価ツールの構築の基礎データにするなど、幅広い活用が見込まれる。本プロジェクトの他のメタデータと組み合わせた研究も考えられるだろう。たとえば、誤用タグと組み合わせて印象評価との関連を分析したり、プロセスデータと組み合わせて執筆に時間がかかった箇所や修

正を繰り返した箇所と作文の内容の一貫性の評価を合わせて分析したりする研究が考えられる。印象評価情報単独での研究だけでなく、他のデータと組み合わせた研究も行われることで、より活用の幅が広がるのではないだろうか。

□6　まとめ

本章では学習者が執筆した作文に印象評価情報を付与する意義を述べたうえで、本プロジェクトが進めている作文の印象評価調査について、パイロット調査の試行錯誤と本調査のデータの概要に分けて論じた。2025年4月現在、本調査によって240本の作文に対する印象評価データが収集されている。このデータはパイロット調査での検討を踏まえて評価者内・評価者間の評価尺度のずれをできるかぎり小さくしたものである。今後この印象評価データが活用され、教育現場の日本語教師や日本語学習者、教育現場の外で日本語の文章を書こうとしている日本語学習者、作文や作文の評価の研究などに広く活用されることを期待して、本章のまとめとしたい。

参考文献

井伊菜穂子（2025）「人間は学習者の作文をどう評価するか」『令和6年度NINJAL日本語教師セミナー「作文教育のための学習者コーパス活用法」発表資料』（2025年3月22日，琉球大学）.

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」李在鎬（編）『文章を科学する』38–57, ひつじ書房.

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―評価プロセスの多様性をとらえることの意義―』ココ出版.

近藤ブラウン妃美（2012）『日本語教師のための評価入門』くろしお出版.

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, 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）『日本語プロフィシェンシー研究の広がり』225–237, ひつじ書房.

田中真理・坪根由香里（2011）「第二言語としての日本語小論文におけるgood writing 評価―そのプロセスと決定要因―」『社会言語科学』14（1）: 210–222.

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, 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.
田中真理・長阪朱美・成田高宏・菅井英明（2009）「第二言語としての日本語ライティング評価ワークショップ―評価基準の検討―」『世界の日本語教育』19: 157–176.
本多由美子（2024）「中国語話者の日本語作文から見る語彙の発達―印象評価データを用いた分析―」『中国語話者のための学習者コーパスを用いた作文の研究と教育』国立国語研究所 機関拠点型基幹研究プロジェクト「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」2024年度研究成果中間報告書.
本多由美子・井伊菜穂子（2025）「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」『国立国語研究所論集』28: 151–166.
McNamara, T. F.（2000）*Language testing*. Oxford University Press.
李在鎬・長谷部陽一郎・村田裕美子（2019）「学習者作文の習熟度に関する自動判定とWebシステムの開発について」李在鎬（編）『ICT×日本語教育』38–53, ひつじ書房.

自動評価ツール（いずれも2025.3.23最終閲覧）

GoodWriting Rater
https://goodwriting.jp/wp/
jWriter 学習者作文指導評価システム
https://jreadability.net/jwriter/

関連URL

「印象評価情報の付与―クラウドソーシングの活用―」印象評価データ
https://doi.org/10.15084/0002000581

（本多 由美子・井伊 菜穂子・石黒 圭）