中級日本語学習者の作文を評価するための汎用性のある評価基準の作成
――JF 日本語教育スタンダードに基づいて――

徐　アルム

Creation of Versatile Writing Assessment Criteria for Intermediate-level Japanese Learners' Essays
– Based on JF Standard for Japanese Language Education –

SEO Areum

Resume

This research reports an investigation into the reliability of the 'Standards for Assessment of Intermediate-level Japanese Learners' Essay' based on relevant recent studies in the field of Japanese language education, the Common European Framework of Reference for Languages: Learning, Teaching, Assessment(CEFR), and the JF standard for Japanese Language Education. The evaluation of 24 intermediate-level Japanese learners' essays from selected the 'JLPTUFS Writing Corpus' was conducted according to the 'Standards for Assessment of Intermediate-level Japanese Learners' Essay'.

The evaluators were four native teachers of Japanese at universities. These evaluators were divided into two groups according to the scoring methods employed: a holistic scoring method (Group A) and a newly developed scoring method (Group B). Correlation analysis, Cohen's kappa, a significance test of correlation coefficient, and intra-class correlation coefficient were conducted using IBM SPSS Software and Microsoft Office Excel.

The results are as follows. (1) Intra-group correlation analysis and Cohen's kappa showed neither significant correlation nor degree of concordance within any evaluator group. Results of inter-group analysis were also unable to be deemed significant or reliable. (2) There was no significant difference between either scoring method. (3) All evaluators provided a similar opinion regarding the advantages of the 'Standards for Assessment of Intermediate-level Japanese Learners' Essay': a clear categorization of evaluation items. On the other hand, 'redundant items' and 'inadequacy and redundancy of evaluation items' were discovered as disadvantages of the rating scale.

Three factors were deemed to contribute to unreliable results: an insufficient number of evaluators, a lack of evaluator training, and the influence of the level of regular classes run by evaluators. In conclusion, creation of a new rating scale that utilizes the advantages observed in the course of this study and includes corrections to the "Standards for Assessment of Intermediate-level Japanese Learners' Essay" should be considered for further research.

目　次

第 1 章　序　論

1. 1　はじめに（研究の背景）

言語教育及び学習において、評価は学習者の言語能力の熟達度や達成度を把握する手段として重要な役割を果たしている。また、学習者を評価することは教授法の評価や改善にも役立つ。しかし、評価方法や評価者により、その結果に差が生じることは避けられない。評価基準の作成やその基準によって行われる評価など、評価に関わる諸過程には必ず人間が関わっているため､主観性を全て排除することは難しい。しかし、訓練された主観を用い、作文評価をより客観的に行うことはできよう。菊池（1987: 90）は、作文評価における主観性介入の問題について以下のように述べている。

> ……（前略）主観が入ることは避けられないし、また、それ自体決して悪いことではない。採点時の身心の状態・気分等によって採点の結果が大きく違うというような主観ではいけないが、一定の意識を備え、かつ、提出された全員の作文に対して一貫性をもって評価できるような―言いかえれ

ば、別の機会に採点しても、同じ作文に対してはほぼ一定の安定した評価を下す結果になることが保証されるような、筋の通った―主観であるならば、それによって評価するところがあってよいであろう。……（後略）

つまり、主観が入っていても、それが一貫性の保たれているものであれば許容されるということである。しかし、いかに一貫性のある主観を保つことができるだろうか。作文評価を構成している諸要素やそれに対する理解から、共通枠組みを抽出し、信頼性と妥当性のある評価基準を構築する必要がある。それによって、一貫性のある主観とともに客観性のある評価が実現できるだろう。

日本語教育分野においても、作文評価における主観性・客観性の問題についての議論がなされており、森田（1980）、斉木他（1988）、菊池（1987）を始め、田中他（1998b）、川上（2005）に至るまで、作文評価基準の作成に関する研究がこれまでにいくつか行われてきたが、汎用性[1]のある作文評価基準の作成とその実践にまでは及んでいない。

国際的な汎用性を持つ評価基準作成の試みもすでに存在している。例えば、Common European Framework of Reference for Languages: Learning, Teaching, Assessment（ヨーロッパ言語共通参照枠、以下、CEFR）や、それを日本語教育の文脈に適用しようとした JF 日本語教育スタンダード（以下､JF スタンダード）がある。しかし、CEFR や JF スタンダードに関する問題が指摘されていないわけではない。「内容における具体性の欠如」､「現場での利用における非効率性」､「使用者により恣意的に用いられてしまう危険性」などの問題が指摘されてきている（国際交流基金 2005: 42）。また、JF スタンダードは、CEFR を基に翻案しているため、以上に述べた CEFR の負の側面まで踏襲してしまっているのが問題点として挙げられる。

したがって、本研究では、以上に述べた CEFR 及び JF スタンダードが持つ短所を克服すると同時に、それらの参照枠の持つメリットを活かすという形で汎用性のある作文評価基準の作成に取り組んだ。それから、全学習レベルにおける評価基準を作成する第一段階として、まず対象を中級日本語学習者に絞った。対象を中級日本語学習者だけに限定したのは、以下の理由による。

小森（2005: 197）によると、中級日本語学習者の作文では、初級の段階に比べ、文型や語彙が増えて表現力が豊かになったり、書き言葉と話し言葉の区別ができるようになったりするなど、文章表現力の向上がみられるという。しかし一方では、文法・語彙・表記などの誤りを直してもなお、読み手に分かりにくい、つまり結束性や卓立性の欠ける作文がよく見られるという。

さらに、CEFR の第 9 章では、初級レベルではまだ学習されてない項目が多いため、現存する殆どの評価尺度を能力記述文などにより解釈した際、否定的表現になりがちであるという弱点があるということが指摘されている。しかし、中級レベルあたりでは、初級レベルとは異なり、規範に準拠する傾向があると述べられている。つまり、初級レベルでは、文章能力より日本語能力の向上や習得に学習の焦点が当てられているため、作文能力を評価することが他の学習レベルに比べて難しいのである。

このようなことから、中級日本語学習者の作文に現れる特徴が評価可能な作文評価基準を作成する必要があると判断した。そのために、先行研究や CEFR、JF スタンダード、そして本研究で用いられた JLPTUFS 作文コーパス、JLC 日本語スタンダーズ（以下、JLC スタンダーズ）の内容や関連項目を参考にしたうえで、新たに項目をまとめ、カテゴリー化する作業を経て出来上がったのが、本研究で提示する「中級日本語作文評価基準」である（2.1 参照）。

続く 1.2 では、第二言語としての日本語教育における作文評価研究についてまとめ、本研究の目的及び、それに伴う研究課題を 1.3 と 1.4 で示す。

1.2　第二言語としての日本語教育における作文評価基準に関する先行研究

ここでは、第二言語としての日本語教育における作文評価基準の先行研究について述べる。

第二言語としての日本語教育における作文評価基準の先行研究として、まず、森田（1980）と斉木他（1988）の研究が挙げられる。森田は、作文の評価方法を大きく「診断的評価」、「形成的評価」、「総括的評価」の3つに分けている。さらに、テストのタイプにおける分類方法を提示し、それらに基づき、「総括的評価」のための評価項目を提案した。しかし、文法的特徴を評価する項目がほとんど立てられていないという短所が見られた。

森田（1980）の研究をより発展させたのが、斉木他（1988）である。斉木他は、「形成的評価」に基づき、森田の基準をより細分化させたが、文法能力を評価する項目が、文章能力を評価する項目より多く設けられたため、文章能力の評価に重点が置かれていないという短所がある。

菊池（1987）は、初級後半から中級までの日本語学習者を対象として、作文を点数で評価する場合の1つの方法を提示した。外国人学習者に日本語の作文を書くための能力を、「趣旨の明確さ」、「内容」、「正確さ」、「表現意欲・積極性」、「表現力・表現の豊かさ」の5つのファクターとして示した。かなり具体的な評価ができる項目構成ではあるが、「減点法による正確さの採点」や「作文の長さによる作文の良し悪しの評価」などが問題点として考えられる。以上に述べた2つの疑問点以外に、川上（2005）も菊池の評価ファクターについて、「作文作成の際に現れる能力を、日本語能力と文章能力に分けて捉えた場合、『主旨の明確さ』のように両方の能力が関わってくるファクターがあると考えられるので、文法的要素を評価する『正確さ』とその内容が重複するのではないか」という疑問を提示した。

一方、田中他（1998）は、汎用性のある作文評価基準の必要性について訴え、日本語教師と一般日本人を対象にし、外国人学習者が書いた作文を評価する際、どのような項目により焦点を当て評価するかについて研究調査を行った。その結果、「正確さ」、「形式・構成」、「内容」、「豊かさ」の4つの因子が抽出された。これらのカテゴリーに属する項目はとても参考になったものの、カテゴリーの振り分けに統一性がないという短所が見られた。

田中他と同様の目的に基づいて行われたもう1つの研究が川上（2005）である。川上は、日本語教師が作文を評価する際、どのような要素により重点を置いているかを把握するために、国内および海外の日本語教育機関で働く現役日本語教師9名を対象に実態調査を行った。初級から上級までの3段階に分けて行われた調査の結果、教師による項目の重視傾向において、レベル別に差が存在するということが分かった。川上の研究は、実態調査という側面では意義があると思われるが、その結果に基づく作文評価基準の作成やその実証までは行われていなかったという限界が見られる。

以上に述べた作文評価基準に関する先行研究に基づき、評価基準項目や分類の仕方について把握することはできたのだが、いずれも基準や方法の提示に留まり、汎用性のある基準の作成やその試みまでは至っていなかった。

1.3　研究の目的

本研究は、以上に述べた日本語教育の分野における作文評価についての先行研究を始め、CEFR及びJFスタンダードに基づいて作成された「中級日本語作文評価基準」の妥当性や信頼性を検証し、汎用性のある作文評価基準の土台を作ることを目的とする。

そのために、現在国内の大学にて日本語を教えている日本語母語話者の日本語教師4名に、筆者が新たにまとめた作文評価基準に沿って、中級日本語学習者の作文を評価してもらう研究調査を行った。評価者は採点方式により、それぞれA（個人別採点方式）とB（新折衷採点方式）の2つのグループに分かれ、作文を評価した[2]。評価する作文データとしては、東京外国語

大学留学生日本語教育センター（以下、JLC）教育研究開発プロジェクトである「JLPTUFS 作文コーパス」から抽出した中級日本語学習者の作文データ 24 個が用いられた。

以上に述べた研究調査の実施結果に基づき、先行研究に提示された作文評価基準項目や CEFR または JF スタンダードが持つ汎用性の検証とともに、グループ別に用いられた異なる採点方式の中で、どちらがより効率性や信頼性に優れているかについて考察する。

1.4　研究課題

本研究における研究課題は以下の通りである。

「中級日本語作文評価基準」による作文評価における評価者間の評価結果について

研究課題 1　グループ A における評価結果の一致度はどの程度か。

研究課題 2　グループ B における評価結果の一致度はどの程度か。

研究課題 3　グループ間の評価結果における一致度はどの程度か。

研究課題 4　本調査にて用いられた 2 つの採点方式のうち、より効率性と信頼性があると思われるのはどちらか。

第 2 章　研究及び分析方法

2.1　「中級日本語作文評価基準」について

ここでは、本研究の主軸であると言える「中級日本語作文評価基準」について概観する。

2.1.1　本研究における「中級日本語学習者」の定義

本研究調査には「JLPTUFS 作文コーパス」のデータが用いられた。「JLPTUFS 作文コーパス」は、東京外国語大学全学日本語プログラム（JLPTUFS）の教育課程の中で、データ提供の同意が得られたものをデータ化し、まとめたものである。JLPTUFS 作文コーパスにおける中級日本語学習者の作文データを使うことにおいては、その背後にある JLC スタンダーズにおける中級日本語学習者の定義、そして、「中級日本語作文評価基準」の作成に大きく参考となった JF スタンダードのレベル分けについて理解しておく必要がある。そこで、本研究における中級日本語学習者の定義の理解を助けるために、5 つの基準――①新日本語能力試験、②旧日本語能力試験、③一般的なレベル分け、④ JLPTUFS のレベル分け、⑤ CEFR に準じた JF スタンダードのレベル分け――を反映させてまとめた学習レベルについて以下に示す。

表 1　本研究における中級日本語学習者の定義

新日本語能力試験	旧日本語能力試験	一般的なレベル分け	東京外国語大学 留学生日本語教育センター 全学日本語プログラム（JLPTUFS）		JF スタンダードのレベル分け
N1	1 級合格	超　級	800	読解・ドラマ・時事・文学・ビジネス日本語・ライティング	C2
	1 級目標	上　級	700	総合　文法・読解・聴解・文章表現・口頭表現・時事日本語	C1
N2	2 級合格	上　級	600	総合　文法・読解・聴解・文章表現・口頭表現	
	2 級目標	中上級	500	総合　文法・読解・聴解・文章表現・口頭表現	B2
N3		中　級	400	総合　文法・読解・聴解・文章表現・口頭表現	B1
N4	3 級合格	初中級	300	総合　文法・読解・聴解・文章・口頭	A2
	3 級目標	初級（後半）	200	初級	
N5	4 級目標	初級（前半）	100	入門	A1

2.1.2 「中級日本語作文評価基準」の作成過程

「中級日本語作文評価基準」は、大きく、文法的項目における作文評価基準表、談話的項目における作文評価基準表、そして、全体的評価の3つで構成されている。ここでは、作文評価基準の作成に主に参考とした先行研究3つのうち、汎用性のある作文評価基準の作成を目指して行われた田中他（1998b）、川上（2005）を始め、英語教育における作文評価における判断特性を文法的項目及び談話的項目に分けて提示したChiang（2003）及び、JFスタンダードが提示したJF Can-do, JLCスタンダーズに提示されている評価項目を以下の方法によりまとめた：

1） 田中他（1998b）、川上（2005）、Chiang（2003）の作文評価項目をまとめ、Chiang（2003）の枠組みに倣い、「談話的項目」及び「文法的項目」に分類する。
2） JFスタンダードのJF Can-do、及びJLCスタンダーズに提示されているレベル別技能一覧を参考にし、中級日本語学習者に求められる作文能力を評価する項目を設ける。
3） Chiang（2003）を参考にし、最後に全体的評価を設ける。

2.1.3 「中級日本語作文評価基準」の提示

以上の過程を経て作成された「中級日本語作文評価基準」は、Chiang（2003）の枠組みを参考にし、大きく「文法的項目における作文評価基準表」、「談話的項目における作文評価基準表」、そして「全体的評価」の3つに分かれている。

2.1.3.1 文法的項目における作文評価基準表

「文法的項目における作文評価基準表」に入る文法的項目を「文法」、「文字・表記」、そして「文字・表記（手書き）」の3つに分類し、まとめた。この評価基準表は、作文の中に現れる項目の「文法的正確さ」にその焦点を当てている。各項目における評価方法としては、5段階評価方法[3]が用いられ、評価者の意見に最も近いところに○をつけてもらう形で評価を行うようにした。また、内容が不十分、もしくは該当する項目がない場合のために、「該当なし」というセルを設けた。

以下に示す表2「文法的項目における作文評価基準表」では、5段階評価の列の代わりに、「中級日本語作文評価基準」の作成の際、主に参考とした3つの先行研究――田中他（1998b）、川上（2005）、Chiang（2003）――及び、JFスタンダード、JLCスタンダーズの内容がいかに反映されたかが示されている。

2.1.3.2 談話的項目における作文評価基準表

談話的項目における作文評価基準表では、作文評価における談話的項目を「内容」、「構成」、「表現」の3つに分類し、まとめた。談話的項目における作文評価は、内容や構成における「一貫性」及び「結束性」、「中級レベルにおける表現能力」に焦点を当てている。文法的項目と同様、5段階評価の選択肢[7]を取り入れ、評価者の意見に最も近いところに○をつけてもらうという形で評価を行うように作成した。内容が不十分、もしくは該当する項目がない場合のために、「該当なし」というセルを設け、○をつけて評価するようにした。

表2「文法的項目における作文評価基準表」と同様、表3「談話的項目における作文評価基準表」でも、5段階評価の列の代わりに、「中級日本語作文評価基準」の作成の際、主に参考とした3つの先行研究及び、JFスタンダード、JLCスタンダーズの内容がいかに反映されたかを示した。

表 2　文法的項目における作文評価基準表

カテゴリー			項　　目	田中	川上	Chiang	JF	JLC
文法的項目	文法 (3)	正確さ (3)	1. 助詞の使い方。	○	○	○		
			2. 動詞や形容詞の活用[4]。	○	○	○		
			3. 作文全体の文法的要素。	○	○	○		○
			3 で見られた要素に✔してください。 （複数可） 自動詞、他動詞　□ 「こ・そ・あ」　□ 副詞　□ 時制　□ 接続語句　□ 主述の対応　□ 語順　□ 数詞　□ その他　（　　）□					
	文字・表記 (4)	正確さ (4)	4. 仮名（ひらがな／カタカナ）表記[5]。	○	○			
			4 で見られた要素に✔してください。 （複数可） ・仮名の字形　□ ・カタカナ語（外来語）表記　□					
			5. 漢字の表記。	○	○			
			6. 単語の表記。（脱字の有無など）		○	○		
			7. 句読法。		○	○		
	文字・表記（手書き[6]） (4)	正確さ (4)	8. 題の位置		○			
			9. 名前の位置		○			
			10. その他（句読点、記号など）		○			
			11. 促音、拗音の位置		○			

2.1.3.3　全体的評価

全体的評価は、作文の総合的側面を評価するために設けられたもので、評価者の「作文全体における印象点」に基づいて行われる。これは、5 段階評価を間隔尺度として考え、点数化した全体得点とは異なる概念である。作文に対する総合的評価は、A（上）、B（中上）、C（中）、D（中下）、E（下）、N/A（該当なし）の中で、評価者の考えが最もよく反映されていると思われるところに○をつけるという形となっている。文法的項目や談話的項目と同様に、該当するところがないと判断された場合、つまり、作文に対する全体的評価ができない場合は、「該当なし」にチェックをする。

2.2　研究の方法

2.2.1　概　要

新たにまとめた作文評価基準である「中級日本語作文評価基準」の汎用性を検証するために、現在国内の大学にて第二言語として日本語を教えている日本語母語話者の日本語教師 4 名に協力してもらい、「中級日本語作文評価基準」に沿って、中級日本語学習者の作文を評価する研究調査を行った。研究調査の実施にあたり、予め作成した「中級日本語作文評価基準」の研究資料を郵便で各協力者宛に送付した。調査実施期間は約 1 ヶ月半で、データとして、JLC の教育研究開発プロジェクトである「JLPTUFS 作文コーパス」から抽出した中級日本語学習者の作文データ 24 個が用いられた。また、グループ A（個人的採点方式）と B（新

折衷採点方式）の 2 つのグループに別れ、グループごとに異なる採点方式が用いられた。研究調査は、以下の手順により行われた。

（1）調査開始の前に、「中級日本語作文評価基準」とともに送付した依頼書と同意書を作成してもらい、事前調査質問紙に答える。

（2）一緒に同封した作文データ（2.2.3 参照）を「中級日本語作文評価基準」に沿って評価する。

（3）作文評価終了の後は、フォローアップ・アンケート及び電話インタビューに答える。

表 3　談話的項目における作文評価基準表

カテゴリー			項　　目	田中	川上	Chiang	JF	JLC
談話的項目	内容（5）		1. タイトルと内容が一致している。		○			
			2. 述べている事柄に魅力がある[8]。	○	○			
			2 で見られた要素に✔してください（複数可） 説明が具体的である　□ 適切な例を提示している　□ 内容の展開が興味深い　□ その他（　　）　□					
			3. 全体として言いたいことが明確である。	○	○	○		
			4. テーマが十分掘り下げられている。	○	○			
			5. 分かりやすく書いてあり、スラスラ読める。	○				
	構成（9）	構成（4）	6. 作文が論理的に構成されている。（起承転結など）	○				
			7. 読者が理解できる、ある程度の長さの文章が書けている。				○	
			8. 標準的な常用形式に沿って書けている。				○	○
			9. 順序立てて並べた書き方になっている。				○	○
		結束性（3）	10. 段落の分け方が適切である。	○	○	○		
			11. 主旨に一貫性がある。	○				
			12. 文と文の繋がりが適切である。	○	○			
		一貫性（2）	13. 文体が統一されている。 （「です・ます」体と「だ・である」体）	○	○			
			14. 話し言葉と書き言葉の使い分けができている。（例：食べちゃった）	○	○			
	表現（12）	豊かさ（1）	15. 言葉や表現が豊かである。	○				○
		表現能力（8）	16. 物事を対比させ、表現する能力が見られる。					○
			17. 事実と考えを分けて書く能力が見られる。					○
			18. 自分の感情を描写する能力が見られる。				○	
			19. 物事の定義がきちんとできている。					○
			20. 物語を書く能力が見られる。				○	
			21. 事実関係を述べ、理由を説明することができている。				○	
			22. 経験や印象を述べられる能力が見られる。				○	
			23. 異文化間の違いに対する認識・配慮があり、それを表現する能力が見られる。				○	
		適切さ（2）	24. 語彙の使い方、選び方が適切である。 （例：×帽子を着る、×薬を食べるなど）	○				
			25. 日本語として意味をなさない文は含まれていない。	○				
		漢字の使用状況（1）	26. 漢字の割合が適度である。 （ひらがなが多すぎないか）	○	○			

2.2.2　協力者

現在日本国内の大学で働く、日本語母語話者の日本語教師 4 名に協力してもらった。いずれも作文指導や評価に興味を持っており、作文や文章表現の指導及び評価の経験がある。研究調査実施の際、4 名の協力者をそれぞれグループ A、B に割り当て[9]、評価者番号を与えた。評価者番号は「グループ名（アルファベット 1 文字)」と「数字（2 桁)」に構成されている。

2.2.3　作文データ：「JLPTUFS 作文コーパス」

「JLPTUFS 作文コーパス」は、JLC の「全学日本語プログラム」（JLPTUFS）の教育課程の学習者が書いた作文の中、執筆者によるデータ提供の同意が得られた作文をデータ化したものである。このプロジェクトは 2009 年から 2010 年にかけて行われたもので、2011 年 3 月に完成・公開された。

その中で、本研究に用いられたデータは、2009 年の総合クラスで宿題として出された 24 人分の作文で、テーマは「留学生のストレス解消法」である。時間や分量には特に制限がなく、作文の際、電子辞書を使っても良いという条件がつけられていた。「JLPTUFS 作文コーパス」の作文データの中で、テーマとインフォーマントの数を考え、このクラスのデータを選んだ。また、JF スタンダードや JLC スタンダーズに提示されている中級日本語学習者の「書くこと」における主な話題や場面として、「良く知っていて、自分の関心事の身近な話題や日常的な事柄」が挙げられるということから、留学生の日常や生活に関係のあるテーマにすることにした。また、インフォーマントも、研究調査の資料として使える程度の数にすべきであると考え、20 人以上の学生で構成されているクラスに絞ってデータを抽出した。その結果、「留学生のストレス解消法」というテーマで書かれた 24 人分の作文が最も適切であると判断し、このデータを本研究調査に用いた。

2.2.4　評価終了後の調査手順：フォローアップ・アンケート及びインタビュー

全ての作文に対する評価の後、各協力者宛に郵送した研究資料に同封されているフォローアップ・アンケートに答えてもらい、研究調査に用いられた資料の全てを返送してもらった。資料が調査実施者宛に届いたことが確認されてから、1 人ずつ、簡単に 5〜10 分程度、電話でインタビューを行った。インタビューは、事前調査質問紙、フォローアップ・アンケート、そして新たに設けたその他の質問に基づいて行われた。全体の流れを以下の表 4 に提示する。

表 4　本研究調査におけるインタビューの手順

順番	カテゴリー	質問内容
1	事前調査質問紙	①作文評価について何を大切にしているか。 ②具体的にどのような方法で評価しているか。 ③今まで作文の授業で使われた教科書は何か。
2	フォローアップ・アンケート	④「中級日本語作文評価基準」について良いと思われた点は何か。 ⑤「中級日本語作文評価基準」で改善が必要であると思われた点は何か。
3	その他の質問	⑥普段の作文評価の際、作文のレベル分けはどのように行っているか（例：3 段階、5 段階)。 ⑦本研究調査における作文評価の全体的評価に際して、A〜E の 5 つのレベル分けに個人的に用いた基準があるか。

2.3　分析方法

2.3.1　データのまとめ①:作文評価データ

収集された作文評価データを評価者別にまとめ、合計4つのエクセルファイルに作文評価データを入力した。全てのデータは、統計的手法により分析された。その詳細については続く2.3.3で述べる。

2.3.2　データのまとめ②：事前調査質問紙、フォローアップ・アンケート、インタビュー

事前調査質問紙とフォローアップ・アンケートも、作文評価データと同じく、評価者別にファイルを作成し、その中に2つのシートを入れ、それぞれアンケートとインタビューのデータを書き込んだ。アンケートの結果は、送ってもらった文面をパソコンで書き写し、データ化する作業を行い、電話インタビューは、重要な内容をメモする形で行われた。本稿では、フォローアップ・アンケートとインタビューの結果のみ3.1.4で述べる。

2.3.3　各研究課題における分析方法及びその意義

以上にまとめた研究調査のデータをいかなる方法で分析したのか、またその意義は何かについて、以下の表5を通じて解説する。

第3章　研究調査データの分析

ここでは、研究調査で得られたデータの分析結果を示す。まず、3.1全体にわたって、統計的手法による作文評価データの分析結果を示す。第3章では、分析の便宜上、文法的項目をG、談話的項目をDと表記し、各項目番号をアルファベットの後に記すことを予め述べておく。

3.1　作文評価データの分析

前述したように、作文評価データの分析の際、主に用いられた統計的手法は、相関分析、2つの相関係数の差の有意差判定公式、カッパ係数、級内相関係数である。ここでは、それらの統計的手法による評価データの分析結果を述べる。

3.1.1　相関分析による作文評価データの分析

2.3.3で述べたように、相関分析は、グループ内の評価者間における評価結果の相関を調べるために用いられた。評価結果を文法的項目、談話的項目、全体的評価、そして、全体得点の4つに分けて分析を行った[12]。各グループ内における分析結果を、3.1.1.1で提示する。また、本研究調査において用いられた採点方式はグループ別に異なったものの、グループ間の評価結果の信頼性において、いかなる差があるかについて

表5　研究課題の内容及びその意義、分析方法のまとめ表

課　題	内　容	意　義	分析方法[10]
研究課題1	グループAにおける評価結果の一致度はどの程度か。	・評価基準の**信頼性**の検証	・相関分析 ・相関係数の有意差判定公式 ・カッパ係数（Cohen's kappa） ・級内相関係数（ICC[11]）
研究課題2	グループBにおける評価結果の一致度はどの程度か。		
研究課題3	グループ間の評価結果における一致度はどの程度か。		
研究課題4	本調査にて用いられた2つの採点方式のうち、より効率性と信頼性があると思われるのはどちらか。	・採点方式の**信頼性**、及び**効率性**の検証	・相関分析 ・カッパ係数（Cohen's kappa） ・インタビュー

検証する必要があるということから、「2 つの相関係数の差の有意差判定公式」を用い、グループ間項目別相関についても分析を行った。その結果は 3.1.1.2 で示す。

3.1.1.1　各グループにおける作文評価データの相関分析

各グループにおける作文評価結果の項目別相関を、文法的項目、談話的項目、全体的評価、そして、全体得点の 4 つに分け、有意であった項目を中心に表 6 にまとめた。分析の際、グループ A では G9、D2、D19、D20 が、グループ B では G9、D18、D19、D20、D23、D26 において相関関係が見られなかったので除外した[13]。

まず、項目別分析結果に基づき、グループ A の評価者間において相関が高かった順に項目を並べると、G10「文法・表記（手書き）：その他（句読点、記号など）」、G5「文字・表記：漢字の表記」、G2「文法：動詞や形容詞の活用」の順となる。この 3 項目は、相関係数が 0.4 以上で、中程度の相関が見られた。また、表には提示していないが、以上に取り上げた項目の他に相関が高かった項目として、G3「文法：作文全体の文法的要素」、G1「文法：助詞の使い方」が挙げられる。全て 0.3 以上の弱い相関が得られた[16]。文法的項目における相関係数の平均は 0.334 [17]であった。

続いて、グループ A の評価者間における談話的項目の分析結果について述べる。26 個の談話的項目のうち相関が高かったのは、D11「構成：主旨に一貫性がある」と D13「構成：文体が統一されている」で、0.7 以上の強い相関が現れた。その次に、中程度の相関が見られたのは、D1「内容：タイトルと内容が一致している」、D6「構成：作文が論理的に構成されている」、D7「構成：読者が理解できる、ある程度の長さの文章が書けている」である。これらの項目は、相関係数が 0.6 以上で、全て 1% 水準で有意であった。

最後に、作文全体における評価の相関は、全体得点（SUM^1）と全体的評価（SUM^2）の 2 つに分けて分析した。全体得点においては 0.6 以上のかなり高い相関が見られたが、全体的評価の場合は、0.4 以上の中程度の相関が現れた。

グループ B の評価者間における項目別相関分析では、文法的項目において、G10「文法・表記（手書き）：その他（句読点、記号など）」、G1「文法：助詞の使い方」、G2「文法：動詞や形容詞の活用」の順で高い相関係

表 6　各グループにおける作文評価の相関分析 [14]

グループ **A**（個人別採点方式）				グループ **B**（新折衷採点方式）			
項　目	相関係数	有意確率	**t** 値	項　目	相関係数	有意確率	**t** 値
G10	0.660	0.000**	4.12	**G10**	0.737	0.000**	5.12
G5	0.606	0.002**	3.57	**G1**	0.725	0.000**	4.94
G2	0.454	0.026*	2.39	**G2**	0.637	0.001**	3.87
D11	0.769	0.000**	5.64	**D13**	0.760	0.000**	5.49
D13	0.727	0.000**	4.97	**D5**	0.608	0.002**	3.59
D1	0.697	0.000**	4.56	**D14**	0.585	0.003**	3.38
D6	0.685	0.000**	4.41	**D22**	0.551	0.005**	3.10
D7	0.618	0.001**	3.69	**D2**	0.519	0.009**	2.84
D9	0.591	0.002**	3.44	**D12**	0.508	0.011*	2.76
D10	0.571	0.004**	3.26	**D6**	0.487	0.016*	2.62
D21	0.528	0.008**	2.92	**D10**	0.485	0.016*	2.60
D8	0.518	0.010**	2.84	**D8**	0.456	0.025*	2.41
D5	0.505	0.012*	2.74	**D4**	0.447	0.028*	2.35
D4	0.472	0.020*	2.51	**D16**	0.437	0.033*	2.28
SUM^1 [15]	0.663	0.001**	3.62	**SUM^1**	-0.009	0.968	-0.04
SUM^2	0.449	0.028*	2.36	**SUM^2**	0.746	0.000*	5.25

**1% 水準で有意である／ *5% 水準で有意である。

数が現れた。表には提示していないが、そのほかに相関が高かった項目として、G4「文字・表記：仮名（ひらがな／カタカナ）表記」、G3「文法：作文全体の文法的要素」、G11「文字・表記：促音、拗音の位置」が挙げられる。これらの項目のうち強い相関が見られたのは、G10とG1であった。全て1%水準で有意であった。文法的項目における相関係数の平均は、0.470であり、グループAよりはある程度大きい相関が現れた。

次に、グループBの評価者間における談話的項目の評価結果ではどのような相関が現れたか。全ての項目のうち相関が高かったのは、D13「構成：文体が統一されている」で、相関係数は0.7以上、1%水準で有意であった。その次に相関が高かったのは、D5「内容：分かりやすく書いてあり、スラスラ読める」、D14「構成：話し言葉と書き言葉の使い分けができている」、D22「表現能力：経験や印象を述べられる能力が見られる」、D2「述べている事柄に魅力がある」であった。これらの項目は、0.5以上の相関係数で、1%水準で有意であった。談話的項目における相関係数の平均は0.402であった。

最後に、作文全体における評価の相関に関しては、グループAと同様、全体得点（SUM[1]）と全体的評価（SUM[2]）の2つに分けて分析を行った。グループBは各項目における相関がかなり高かったものの、全体得点の結果においては、-0.009の負の相関が現れた。しかし、全体的評価においては0.7以上の相関が見られ、グループAと同じく、全体得点と全体的評価の相関において相違点が現れた。グループA及びBにおける全体得点と全体的評価の相関の差についての考察は、4.1の総合考察及び結論で述べる。

3.1.1.2　グループ間における作文評価データの相関分析

ここでは、グループAとグループBにおける作文評価の信頼性に差が存在するかについて調べる。そのために、「2つのグループにおける相関係数に差がない」を帰無仮説（H_0）に設定し、「2つのグループにおける相関係数に差がある」を対立仮説（H_1）として立てた。分析の際は、Chiang（2003）に倣い、値をフィッシャーのZ変換により変えた後、t検定を行った。その結果を表7に示す。

表7　2つのグループの作文評価データの相関係数における有意差検定結果
（フィッシャーZ変換後）

項目	グループA	グループB	有意確率	t値	項目	グループA	グループB	有意確率	t値
G1	0.348	0.919	0.066	-1.848	**D1**	0.861	0.159	0.023*	2.275
G2	0.490	0.753	0.395	-0.853	**D3**	0.043	0.292	0.424	-0.808
G3	0.355	0.688	0.285	-1.078	**D4**	0.513	0.481	0.920	0.102
G4	0.280	0.734	0.142	-1.472	**D5**	0.556	0.706	0.631	-0.485
G5	0.703	0.298	0.190	1.310	**D6**	0.838	0.532	0.322	0.992
G6	0.307	0.318	0.976	-0.034	**D7**	0.722	0.238	0.119	1.568
G7	0.224	0.331	0.734	-0.349	**D8**	0.574	0.493	0.795	0.262
G8	-0.063	-0.087	0.944	0.079	**D9**	0.679	0.197	0.119	1.563
G10	0.793	0.945	0.624	-0.492	**D10**	0.649	0.529	0.704	0.389
G11	0.224	0.621	0.201	-1.287	**D11**	1.018	0.291	0.019	2.355
SUM[1]	0.798	-0.009	0.009**	2.615	**D12**	0.433	0.559	0.689	-0.409
SUM[2]	0.483	0.964	-1.557	0.120	**D13**	0.922	0.996	0.818	-0.240
					D14	0.494	0.669	0.569	-0.570
					D15	0.439	0.215	0.472	0.727
					D16	0.131	0.468	0.276	-1.094
					D17	0.369	0.239	0.674	0.420
					D21	0.587	0.418	0.589	0.550
					D22	0.286	0.620	0.280	-1.083
					D24	0.181	0.203	0.944	-0.072
					D25	0.168	0.422	0.412	-0.826

**1%水準で有意である／*5%水準で有意である。

分析の結果、項目 D1「タイトルと内容が一致している」と全体得点（SUM[1]）における相関係数以外に、グループ間の相関係数に有意な差がなかったということが分かった。つまり、これら 2 つの項目に関しては、グループ A の方がグループ B より信頼性があるという結果が現れ、2 つの項目に限って採点方式による信頼性の違いが見られたということが分かった。しかし、他の項目においては有意な結果が現れなかったため、2 つのグループに用いられた個人別採点方式と新折衷採点方式における大小関係の相関に関しては、それほど大きい差はなかったと言えよう。

3.1.2　カッパ係数・級内相関係数によるデータの一致度分析

以上に、相関分析による分析結果について述べた。しかし、相関係数は 2 つのデータの平均からの差における一致性、つまり大小関係の一致性を示す指標であるため、2 人の評価者のデータにおける一致度を求めるには限界がある。したがって、データの値そのものが一致しているかを確かめるために、各グループのデータをカッパ係数により分析した。また、異なる条件が設定されているグループ A と B の間における比較に、カッパ係数を用いて分析することは困難であるということから、複数の検者（評価者）によって複数の被験者（学生の作文データ）を評価する場合における信頼性（Inter-rater Reliability）の指標を求める分析手法である級内相関係数[18]を用いた。それぞれの統計的手法による分析結果を、3.1.2.1 と 3.1.2.2 に示す。

3.1.2.1　カッパ係数による項目別一致度分析（グループ別）

カッパ係数は 0 から 1 までの値をとり、値が 1 に近いほど一致度が高いことを意味する。Cohen（1960）によると、カッパ係数が 1 である場合は「完全一致」、0 である場合は「一致してない」と解釈することができると述べられている。また、負のカッパ係数は、「2 人の被験者における一致度が偶然一致する確率より低い」ことを意味する。一般に、カッパ係数は、値が 0.6～0.8 の場合は実質的に一致しているとみなされ、0.8～1 であれば、ほぼ完全に一致しているとみなす傾向が多い。以下にカッパ係数より分析した、2 人の評価者間における項目別一致度をグループ別に示す。

表 8　カッパ係数よる項目別一致度分析（グループ別）[19]

項目	グループ A	グループ B	項目	グループ A	グループ B
G1	0.02	0.188	**D9**	0.125	-0.005
G2	-0.016	-0.002	**D10**	-0.084	0.067
G3	0.012	0.196	**D11**	-0.143	-0.016
G4	0	0.259	**D12**	0.014	0.176
G5	-0.077	-0.208	**D13**	0.186	0.554
G6	0.013	0.036	**D14**	0.105	0.429
G7	-0.079	-1	**D15**	-0.027	-0.027
G8	-0.029	-0.034	**D16**	-0.065	-0.015
G10	0.029	0.331	**D17**	-0.084	0.033
G11	-0.023	0.226	**D18**	-0.025	-
D1	-0.063	0.049	**D21**	-0.043	0.174
D2	-	-0.145	**D22**	-0.059	-0.033
D3	0.081	0.029	**D23**	-0.011	-
D4	-0.153	0.069	**D24**	-	0.168
D5	-0.063	0.271	**D25**	-0.029	0.107
D6	-0.007	0.054	**D26**	-0.078	-
D7	-0.079	0.035	**SUM[1]**	-0.003	-0.019
D8	-0.014	0.007	**SUM[2]**	0.022	0.286

グループAにおけるカッパ係数は、負の値が半分以上であった。つまり、前述したように、「偶然一致する確率より低い」ということである。3.1.1.1で述べた相関係数による分析結果で高い相関が現れた項目においても、一致度が低いという結果が出た。これは、グループAの評価者間におけるデータの値そのものは完全に一致していないが、それらのデータにおける相関は中程度のものであると解釈できる。また、普段大学1年生の文章表現のクラスを担当している評価者A02は、本研究に用いられた中級日本語学習者の作文データの評価において、他の評価者より厳しく評価する傾向を見せた。それも、グループAの低いカッパ係数の値に影響したもう1つの要因として考えられる。

一方、グループBは、グループAに比べ、負の値が少なく、数値も比較的高かったものの、有意な結果の判断基準となる0.6以上の値は見られなかった。しかし、2つのグループにおいて、カッパ係数の値がいちばん高かった項目はD13「構成：文体が統一されている」であったことがわかった。また、グループBもグループAと同様、相関分析の分析結果においては中程度の相関が現れたが、個々のデータにおける一致度、つまり単一測定値の完全一致度という観点からはかなり低い一致度が出たということが分かった。

3.1.2.2　級内相関係数による項目別一致度分析（評価者別）

ここでは、級内相関係数を用い、評価者全体における項目別一致度について分析した結果を述べる。結果を提示するにあたり、統計的理解を助けるために表に提示されている用語について定義しておく。「単一測定値」は、個々のデータの値に対する信頼性を意味し、「平均測定値」は、4名の評価者における平均値データの信頼性を表す。級内相関係数の判定基準として、Landis（1977）、栗原他（1993）、Portney（1993）が提示したものがあるが、その内容から判断して、級内相関係数の値が0.7以上であれば、信頼性は良好であると判断して良いだろう。そして、分析の際、「クロンバックのアルファ係数（Cronbach's α Coefficient、以下クロンバックα係数）[20]」が分析結果の解釈における参考のために用いられた。George（2003）によると、クロンバックα係数の値は、0.8以上であると信頼性が良いと述べられている。

まず、全評価者における単一測定値の級内相関係数から述べる。カッパ係数による分析結果からも明らかになったように、データそのものの一致度はグループ内、及びグループ間において非常に低いということが分かる。一方、各項目のデータの平均値における評価者間信頼性に関しては、単一測定値における分析結果より高い数値が現れた。これは、相関分析による分析結果と結びつけて解釈することができる。3.1.1.1から3.1.1.2にわたって行われた相関分析では、各グループ内評価者間のデータにおいて中程度の相関が現れた。また3.1.1.3で述べた、グループ間における相関係数の有意差検定の結果においても、2つのグループの間にそれほど大きい差はないということがわかった。

以上を踏まえると、個々人のデータにおける完全一致度は低いものの、グループ内及びグループ間の平均による相関や一致度は高いとまとめることができる。それは、クロンバックα係数からも把握できる。α係数はある項目間における内的蓋然性を調べるためのものである。表9に示したクロンバックα係数の平均は0.640[21]で、前述した判断指標の値である0.8には及ばない数値ではあるが、許容される（Acceptable）程度である（Landis 1977）。また、一部の項目や順序尺度である全体的評価においては、0.7〜0.9の高い値が見られ、評価者間における信頼性はある程度保たれていると解釈することができよう。

3.1.3　各グループにおける作文評価時間の分析

第2章の研究方法で述べたように、本研究調査では、作文評価において2つの採点方式を用い、どちらがより信頼性及び効率性があるかについて調べた。調査対

表 9　級内相関係数及びクロンバック α 係数による項目別一致度分析（評価者別）

項　目	単一測定値	平均測定値	有意確率	**Cronbach α 値**
G1	0.159	0.431	0	0.705
G2	0.177	0.463	0	0.705
G3	0.211	0.518	0	0.727
G4	0.266	0.592	0	0.784
G5	0.234	0.55	0	0.819
G6	0.083	0.265	0	0.561
G7	0.121	0.355	0.003	0.581
G8	0.222	0.533	0.007	0.539
G9	0.243	0.562	0.003	0.574
G10	0.55	0.83	0	0.903
G11	0.19	0.484	0	0.687
D1	0.149	0.411	0	0.719
D2	0.022	0.081	0.03	0.451
D3	0.166	0.444	0.001	0.628
D4	0.258	0.582	0	0.777
D5	0.275	0.603	0	0.777
D6	0.402	0.729	0	0.829
D7	0.205	0.508	0	0.75
D8	0.201	0.502	0	0.671
D9	0.262	0.587	0	0.765
D10	0.351	0.684	0	0.754
D11	0.389	0.718	0	0.844
D12	0.326	0.659	0	0.791
D13	0.626	0.87	0	0.923
D14	0.379	0.709	0	0.811
D15	0.059	0.199	0.016	0.492
D16	0.223	0.535	0	0.687
D17	0.191	0.485	0.001	0.636
D18	0.008	0.03	0.175	0.254
D19	0.002	0.01	0.353	0.103
D20	0.001	0.005	0.404	0.062
D21	0.23	0.544	0	0.684
D22	0.099	0.304	0	0.657
D23	0.005	0.02	0.266	0.174
D24	0.183	0.472	0.002	0.605
D25	0.207	0.51	0.001	0.643
D26	0.114	0.34	0.002	0.593
SUM[1]	0.048	0.167	0.054	0.4
SUM[2]	0.382	0.712	0	0.871

象である評価者の数が少なかったため、本調査で得られた結果を一般化するのは容易ではない。しかし、三谷他（2004）に取り上げられた個人的採点方式及び折衷採点方式をさらに発展させた「新折衷採点方式」による作文評価の結果がどうであったかについて理解することにその意義があると考えられる。以下に、評価者及びグループ別の作文評価時間の平均データを示す。

評価者 A01 は、文法的項目および談話的項目の評価にかかった時間は平均 5 分程度であり、同じ時間間隔で評価を行う傾向が見られた。一方、A02 は、文法的項目における評価で平均 18.88 分、談話的項目における評価では平均 17.29 分がかかり、同じグループである A01 に比べて評価時間がやや長かったことが分かった。

グループ B の評価者においては、評価時間にそれほど大きい差は見られなかった。B02 のほうが B1 よ

表 10　評価者およびグループ別の作文評価時間の平均［単位：分］

評価者	平均時間		グループ	平均時間	
	G	D		G	D
A01	5.00	10.00	A	11.93	13.64
A02	18.88	17.29			
B01	5.70	5.17	B	7.22	6.09
B02	8.75	7.02			

り評価における時間が少し長かったものの、それぞれの領域に対して、作文評価における各評価者の評価時間にはバランスが保たれていることが分かった。三谷他（2004）によると、項目別採点方式は個人別採点方式より採点に時間がかかると述べられていたが、グループ A と B における文法的項目の作文評価時間データからは、採点方式による評価時間の違いは特定されなかった。

また、A01 と B01 は文法的項目における評価時間において、平均時間が約 5 分程度で殆ど同じであったが、A02 と B02 における文法的項目の評価時間は他の 2 名の評価者より長かった。先行研究では、採点方式の影響により作文評価にかかる時間に違いが生じるということについて指摘されているが、本研究における文法的項目の評価に採点方式の影響は殆どなかったといっても良いだろう。以上のことから、本調査において、作文評価時間に差が生じる主な原因は、採点方式の問題ではないことが分かった。

3.1.4　フォローアップ・アンケートおよびインタビューの分析

研究調査後行ったフォローアップ・アンケートで、4 名の評価者は、「中級日本語作文評価基準」のメリットについて「評価項目が簡潔に分類されていること」という共通の意見を述べた。その他に、「作文評価を数値に変換することで客観性が保たれること（A01）」、「表現に関する評価項目が多いので、学生の表現能力が評価しやすい（B01）」などの意見もあった。一方、改善が必要な点について、共通する意見はなかったが、「項目の数が多すぎる（A02）」、「評価項目の重複及び不備（A01、B01、B02）」に対する意見が多かった。

一方、各評価者は普段の評価において以下のような評価方法を取り入れていることが分かった：

1）内容、文法、印象点に基づいて 4 段階に分けて評価を行う（A01）。
2）3 段階に分けて全体評価を行う（A02）。
3）文法・表記及び談話に重点を置き、クラスの人数やレベルに応じて 3 段階または 5 段階で評価し、文法や表現の間違いは減点法を用いる（B01）。
4）談話的項目に重点をおいて、3 段階に分けて全体評価を行い、減点法を用い文法や表現を評価する（B02）。

3.3　分析結果のまとめ

以上に述べた分析結果を、以下に箇条書きでまとめる。

（1）各グループ内評価者間における評価結果の相関及び一致度、そしてグループ間における評価結果の相関及び一致度において、汎用性や信頼性が判断できる有意な結果は見られなかった。また、CEFR や JF スタンダードの Can-do に基づいて設けられた談話的項目カテゴリーの「表現能力」に属するいくつかの項目には弁別力がないということが分かった。
（2）採点方式による作文評価結果の信頼性や妥当性の違いは、本研究調査において現れなかった。
（3）「中級日本語作文評価基準」の良い点として、4

名の評価者は「評価項目が簡潔に分類されていること」を共通の意見を述べた。一方、改善点における共通の意見はなかったが、「項目の数」、「評価項目の重複及び不備」の 2 点が挙げられた。

また、上記の分析結果の他に、疑問点として残ったものが 2 点あった。第一は、分析の際、相関を求めることのできない項目がいくつかあったこと、第二の疑問点は、全体得点における相関と、全体的評価における相関に相違が現れたことである。これらの疑問点については、総合考察で述べる。

第 4 章　結論

4. 1　総合的な考察及び結論

ここでは、上記の分析結果に基づき、1）本研究調査で有意な結果が導き出されなかった原因、2）「中級日本語作文評価基準」の今後の方向性、3）作文評価における主観性および客観性という 3 つの観点から考察を行う。

4. 1. 1　有意な結果が導き出されなかった原因

有意な結果が現れなかった原因の 1 つとして「充分でない評価者の数」が考えられる。本調査における評価者の数は 4 名で、その中でも採点方式により 2 名ずつ 2 つのグループに分かれた。ある基準における妥当性や信頼性を調べるには、少ない数であった。この数は、採点方式を比較するという側面においても、適切ではなかったと思われる。何故かというと、データに有意な結果が現れたとしても、それが「採点方式」の違いによるものか、「評価者個人の評価スタイル」の違いによるものか判断し難いからである。

次に、「評価者における事前トレーニングの有無」が挙げられる。評価者も、本研究調査に存在するいくつかの変数の 1 つであるため、作文評価過程に対して精巧なトレーニングを行うのは、ある意味変数を操作することとなる。また、それにより予測できない結果が導かれてしまう恐れがある。しかし、作文評価の過程におけるトレーニングの代わりに、作文評価項目における詳しい説明を伴うトレーニングを個人的に行うことは必要であると思われる。これは、先行研究である Chiang（2003）にも指摘されていることであるため、調査を開始する前にマニュアルの送付とともにそれについての説明を行った。しかし、個々の評価項目が何を評価しようとするかについては、追加説明を詳細に行わなかったため、相関や一致度において有意ではない結果が現れたと考えられる。

以上に述べた 2 つの原因以外に、作文評価に影響を与えるもう 1 つの要因として「評価者が普段担当している授業による影響」がある。中級レベルの学習者だけでなく、それよりさらに上のレベルの学習者が書いた作文によく接する評価者の場合、中級日本語学習者の作文をそれらのレベルと比較してしまい、作文評価により厳しくなる可能性がある。実際、評価者 A02 は、普段担当している授業のレベルが上級、超級に該当するため、中級日本語学習者の作文を評価する際、かなり厳しく評価する傾向を見せていた。以上のことを踏まえ、ある特定のレベルにおける作文評価基準に沿って評価基準を行うにあたり、評価者が現在担当している授業や、これまでの日本語教授歴などをより綿密に検討する必要があると思われる。

4. 1. 2　「中級日本語作文評価基準」の今後の方向性―中級学習者の表現能力項目を中心に―

3.3 で述べた本調査の分析結果のほかに、「項目別相関を求めることのできなかった項目」がいくつかあったことを第一の疑問点として取り上げた。その項目は、G9「名前の位置」、D18「自分の感情を描写する能力が見られる」、D19「物事の定義がきちんとできている」、D20「物語を書く能力が見られる」、D23「異文化間の違いに対する認識・配慮があり、それを表現する能力が見られる」、D26「漢字の割合が適度である」

の 6 つである。上記の項目において、何人かの評価者は全ての作文において同じ評価点をつけた。つまり、データにばらつきがなかったため、相関を求めることができなかったのである。

これらの項目は、作文評価において弁別力がなかった項目としてみなしても良いだろう。項目の弁別力に関連する意見として、評価者 A01 のコメントが参考できる。A01 は、G9「名前の位置」が属するカテゴリーである「文字・表記（手書き）」について、「『文字・表記（手書き）』の項目は、一度習得したらあまり間違いが発見されないため、基準に入れなくても良い」という意見を述べた。G9 においては 2 つのグループとも相関を求めることができなかったため、このカテゴリーに関しては今後修正の必要があると考えられる。

また、以上に述べた項目以外のもののうち、共通のカテゴリーを持つのは D18、D19、D20、D23 で、全て「表現能力」に該当する。これらは、B1 レベルにおける CEFR 及び JF スタンダードの「書くこと」の Can-do を抽出し、評価項目として設けられたものである。本研究調査における作文評価で、「表現能力」の評価項目を評価に取り入れた評価者（A01、B01）とそうでない者（A02、B02）に両分されていた。

このように違う傾向が現れたのは、本研究調査における作文データの位置づけが各評価者において異なったということから起因する。つまり、フォローアップ・アンケートやインタビューで明らかになった「各評価者における内在的な評価基準の相違」によるものであると言えよう。次の原因として、「CEFR や JF スタンダードが持つ特別な性質」が考えられる。本調査において、CEFR や JF スタンダードの Can-do を参考にして設けた「表現能力」カテゴリーは、評価者により恣意的に捉えられ、評価する教師もいればそうでない教師もいた。これは、冒頭で述べた CEFR の負の側面が現れたとも言えよう。

以上を踏まえると、今後、CEFR や JF スタンダードにより設けられた「中級学習者の表現能力」項目を作文評価に効率よく活かすためには、共通参照枠や Can-do の特性を考え、評価の際、評価者が選択的に取り入れるという指示を書き加えるべきであると考えられる。

4.1.3　作文評価における主観性および客観性

「相関を求めることのできない項目」のほか、第二の疑問点として、「全体得点における相関と、全体的評価における相関の相違」が挙げられる。これらの項目における相関や一致度の分析結果において、どちらのグループも違いを見せていた。本研究調査では、作文評価の持つ「主観性」と、作文評価が追求する「客観性」の 2 つの領域を取り入れるということから、文法的項目や談話的項目においては客観性を保ち、全体的評価では作文全体に対する主観的な意見に基づいて評価を行うように取り組んだ。

つまり、文法的項目と談話的項目を得点化して、作文に対する全体的評価を行うというより、作文における「主観的印象」に焦点を当て、全体的評価をするようにしたため、このような相違が生じたのは当然のことであると考えられる。実際、評価者の中で、どちらに重点をおいて評価すればよいか分からないという意見を提示した評価者もいた。しかし、それは、評価者が重視している領域や作文評価の目的に応じて変わりうるものであると思われる。

以上を踏まえ、本研究において把握された「中級日本語作文評価基準」の問題点を修正し、新たに評価基準を作成する際は、利用者が作文の目的に応じて、評価項目リストにある項目を柔軟に取り入れ、作文評価を行うことができるように工夫する必要があるということが分かった。

4.2　今後の課題

以上に、作文評価データの分析結果及び、その結果の中で疑問に残った点に基づいて考察を行った。本研究調査における最も大きな問題点として、調査の実施

にあたり、「中級日本語作文評価基準」を提示するとともに、それらに関する追加的な説明や工夫が適切に行われていなかったことが挙げられる。また、JF スタンダードの枠組みや、その内容を反映することにおいて、JF スタンダードの持つ特性をより考慮すべきであったことも今後の課題として考えられる。すなわち、JF スタンダードに提示されている Can-do を活かして設けた評価項目を、作文評価の際、評価する作文の位置づけや目的に応じて教師が選択的に取り入れることができるように工夫する必要がある。

以上を踏まえ、本研究調査により明らかになった「中級日本語作文評価基準」の長所は活かす一方、現れた問題点を克服し、さらに発展された評価基準を立ち上げることを今後の目標とする。また、本研究を踏み台に、長期的な計画として、本研究における評価の枠組みの更なる実践、及び初級や上級日本語学習者における作文評価基準の作成などを今後の課題として考えている。以上に述べたことが、今後の日本語教育における作文指導や評価に貢献することを望む。

注

1 本稿で定義する汎用性とは、「あらゆる場面において，妥当性・信頼性が確保されていること」を指す。
2 本研究では、三谷他（2004）に提示されている「個人別採点方式」、「項目別採点方式」、そして「折衷採点方式」を参考にし、「個人別採点方式」及び「新折衷採点方式」という 2 つの採点方式を取り入れた。まず、「個人別採点方式」は、作文ごとに全ての観点項目を採点していくやり方で、「新折衷採点方式」は、「文法的項目」の評価には、観点項目ごとに全作文を採点していくやり方である「項目別採点方式」を用いて評価する一方、「談話的項目」の評価には「個人別採点方式」を取り入れるものである。
3 文法的項目における 5 段階評価の選択肢は、「5　とてもよくできている、4　まあまあできている、3　ふつう、2　あまりできていない、1　全然できていない」である。
4 項目 3 の評価は、①「作文全体の文法的要素」における正確さについて評価し、②現れた文法項目をチェックする。該当する項目がない場合は、「その他」に記述する。
5 項目 4 の評価も、項目 3 と同じく、①作文全体における仮名表記の正確さについて評価し、②該当する項目にチェックする。
6 作文が手書きである場合は、「文字・表記（手書き）」の評価項目も含める。
7 談話的項目における 5 段階評価の選択肢は、「5　とてもそう思う、4　そう思う、3　わからない、2　そう思わない、1　全くそう思わない」である。
8 項目 2 の評価は、①作文全体において評価し、②そこに現れた下位項目をチェックする。該当する項目がない場合は、「その他」に記述する。
9 グループの割り当てはランダムに行われた。
10 以上に述べた全ての統計的手法による分析は、Microsoft Office Excel（以下、エクセル）及び IBM SPSS Software（以下、SPSS）で行われた。
11 統計学分野では項目特性曲線（Item Characteristic Curve）の略称を ICC として用いるため，級内相関係数を ICC と略すのは一般的ではないゆえに混乱を招く。論文などで利用する際は，まず「級内相関係数」と断ってその後に略す必要がある（対馬永輝 http://www.hs.hirosaki-u.ac.jp/~pteiki/research/stat/）。
12 「全体的評価」は、作文全体における印象点に基づいて作文の総合的側面を評価するものである。一方、「全体得点」は、文法的項目および談話的項目における 5 段階評価を間隔尺度として考え、点数化したものである。
13 評価データにばらつきがなかったため、相関係数を求めることができなかった。
14 統計量の本来尺度は順序尺度（Ordinal Scale）であるが、便宜上間隔尺度（Interval Scale）に変更して分析を行った。また、各領域において相関係数が高い順に並べた。
15 第 3 章の統計的分析結果において、SUM^1 は、5 段階評価を間隔尺度として考え、点数化した「全体得点」を、SUM^2 は、順序尺度である A~E による「全体的評価」を意味する。
16 石川他（2010:86）によると、一般的に、相関係数の絶対値が 0.7 より大きければ「強い相関」、0.4 より大きければ「中程度の相関」、0.2 より大きければ「弱い相関」、そして 0.2 以下の場合は「相関がない」

とみなして良いと述べられているが、これはあくまでも一般的な傾向にすぎないものであり、絶対的基準ではない。

17　単純平均により求められた値である。今後述べる全ての相関係数の平均は単純平均により出されたことを予め述べておく。

18　Bartko（1966）によると、級内相関係数には一元配置変量（One-way Classification）、二元配置変量（Two-way Random Model）、二元配置混合（Two-way Mixed Model）の 3 つのモデルがあるという。本研究調査では、以上の 3 つの中で「二元配置変量」モデル（完全一致）を選択し、分析を行った。

19　2 つのグループにおいてカッパ係数を求めることのできなかった項目は除外した。

20　クロンバック α 係数は、信頼性係数の一種で、複数の質問項目を加算して何らかの概念を測定する尺度を構する場合に、それらの質問項目間に内的整合性（Internal Consistency）があるかどうかを調べるための指標である。1951 年にクロンバックにより「アルファ」と命名された（Cronbach 1951）。

21　文法的項目と談話的項目における α 係数単純平均による値である。

参考文献

◎日本語

阿野幸一・ベッツ・ロバート・福田浩子・永井典子・岡山陽子・佐々木美帆・上田敦子（2007）「ヨーロッパ言語共通参照枠に基づく英語能力記述尺度：茨城大学総合英語プログラムにおけるケーススタディ」『人文コミュニケーション学科論集』2, pp. 1-18，茨城大学人文学部.

石川慎一郎・前田忠彦・山崎誠（2010）『言語研究のための統計入門』くろしお出版.

川上麻理（2005）「汎用性のある作文評価基準の提案を目指した評価項目の検討―日本語教師を対象とした実態調査を通して」『ICU 日本語教育研究 2』pp. 23-33.

菊池康人（1987）「作文の評価方法についての一試案」『日本語教育』63, pp. 87-104，日本語教育学会.

国際文化フォーラム（中野佳代子ほか）（2007）『高等学校の中国語と韓国朝鮮語：学習のめやす（試行版）』財団法人国際文化フォーラム.

国際交流基金（2005）『ヨーロッパにおける日本語教育と Common European Framework of Reference for Languages』

国立交流基金（2009）『JF 日本語教育スタンダード試行版 2010』

_____（2010a）『JF 日本語教育スタンダード 2010』

_____（2010b）『JF 日本語教育スタンダード利用者ガイドブック 2010』

駒田朋子・安井朱美・山田真理（2008）「中上級学習者の作文を評価する―アカデミック・ライティング評価基準をどう使えばいいか」『南山大学国際教育センター紀要』9, pp. 71-85, 南山大学国際教育センター.

小森万里（2005）「中級作文におけるわかりにくさの要因―結束性、卓立性を支える要素をめぐって―」『山口幸二教授退職記念論集「ことばとそのひろがり（4）」』pp. 197-216，立命館大学法学会.

倉澤永吉（1987）「作文の教師」p. 177，国土社.

栗原洋一・斉藤俊弘他（1993）「検者内および検者間の Reliability（再現性、信頼性）の検討」『呼と循ゼミナール』41, pp. 945-952，医学書院.

三谷閑子・村上京子・小室輝代（2004）「作文の評価手順が評価に及ぼす影響について --analytic scoring の採点に関して」『言語と文化』5，pp. 1-16，名古屋大学大学院国際言語文化研究科日本言語文化専攻.

森田冨美子（1980）「作文の評価」『日本語教育』43，pp. 17-33，日本語教育学会.

斎木ゆかり・照木ミドリ・川幡愛恵美（1988）「作文評価の標準化のために」『東海大学紀要』8, pp. 53-69, 東海大学.

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，pp. 1-12，日本語教育学会.

東京外国語大学留学生日本語教育センター（2011）「JLPTUFS 作文コーパス」

東京外国語大学留学生日本語教育センター（2011）「JLC 日本語スタンダーズ 2011 改訂版」

吉島茂・大橋理枝（訳、編）（2004）『外国語の学習、教授、評価のためのヨーロッパ共通参照枠』朝日出版社.

吉島茂（2008）「セミナー：文化と言語の多様性の中の Common European Framework of Reference for Languages: Learning, teaching, assessment（CEFR）―それは基準か？」『明海大学大学院応用言語学研究』10, pp. 33-43, 明海大学大学院応用言語学研究科紀要編集委員会.

◎英語

Bartko, J.J.(1966) The Intraclass Correlation Coefficient as a Measure of Reliability: *Psychological Reports*, 19, pp. 3-11.

Chiang, Steve(2003) The Importance of Cohesive Conditions to Perceptions of Writing Quality at the Early Stages of Foreign Language Learning, *System*, 31, pp. 471-484.

Council of Europe(2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment,* Cambridge: Cambridge University Press.

Cronbach, L. J. (1951) Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16(3), pp. 297-334.

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales, *In Educational and Psychological Measurement*, XX(1), pp. 37-46.

Cohen, J. (1968) Weighted Kappa; Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit, *Psychological Bulletin.*

Cramer, E.M.(1985) Multicollinearity, in Kotz, S&Johnson, N.L. ed, *Encyclopedia of Statistical Science*, 5, pp. 639-643, John Wiley.

George, Darren and Mallery, Paul (2003) *SPSS for Windows Step by Step: A Simple Guide and Reference*, 11.0 update (4th ed.), Boston: Allyn & Bacon.

Landis, J.R. and Koch, G..G.(1977) The Measurement of Observer Agreement for Categorical Data, *Biometrics*, 33, pp. 159-174.

Jacobs, H.L., Zinkgraf, S.A., Wormuth D.R., Hartfiel, V.F. and Hughey, J.B. (1981) *Testing ESL Compostion; A Practical Approach,* Rowley, MA: Newbury House.

Portney, L.G. and Watkins, M.P.(1993) *Foundations of Clinical Research: Applications to Practice*, pp. 515-516, Appleton & Lange, USA