多言語母語の日本語学習者横断コーパス(I-JAS)を基にした日本語のストーリーライティング評価基準の開発と評価[1]

大和 祐子[2]
玉岡 賀津雄[3]
斉藤 信浩[4]

DOI: 10.18999/stul.33.55

要約:本研究では，日本語教育におけるストーリーライティング(SW)能力(あるいは説明力)の育成において必要とされる評価基準作成のための手順と有用性を検討した。まず，先行研究を参考にして，日本語における SW 能力について，「伝達性」，「結束性」，「表現の豊かさ」，「表現の正確さ」の 4 つの特性を設けた。そして，それぞれの特性について，6 つのレベルからなる評価基準を設定した。次に，開発した評価基準の有用性を検討するために，「多言語母語の日本語学習者横断コーパス(I-JAS)」の SW 課題を用いて評価した。その結果，2 種類の SW 課題についての 2 名の評価者の相関(評価者間信頼性)は高かった(SW1 が r= 0.82，SW2 が r=0.84)。また，日本語能力テストと SW の評価得点との相関も，0.41 から 0.55 の範囲(表 5 を参照)で比較的高かった。さらに，2 種類の SW 課題の評価の相関は，0.77 から 0.87 で非常に高かった(表 6 を参照)。以上の結果から，本研究で開発した SW の評価基準は，日本語教育において有用であることが示された。

キーワード:ストーリーライティング(SW) 評価基準 信頼性 多言語母語の日本語学習者横断コーパス(I-JAS) 説明力

---

1 English Title: Developing evaluation criteria for Japanese narrative abilities using the corpus data of the International Corpus Japanese as a Second Language

2 YAMATO, Yuko (Associate professor, Center for Japanese Language and Culture, Osaka University, Osaka, Japan, E-mail: y.yamato@cjlc.osaka-u.ac.jp)
3 TAMAOKA, Katsuo (Professor, Graduate School of Languages and Cultures, Nagoya University, Nagoya, Japan, E-mail: ktamaoka@gc4.so-net.ne.jp)
4 SAITO, Nobuhiro (Associate professor, International Student Center, Kyushu University, Fukuoka, Japan, E-mail: saito.nobuhiro.489@m.kyushu-u.ac.jp)

1. はじめに

日常生活において，自分が経験したことや見聞きしたことを文章にして伝えなければならない場面は多い。さらに，近年，e-mail，LINE，WeChat，Skype などの普及により，文字を介したコミュニケーション場面で，母語はもちろん母語以外の言語を使うことも急速に増えている。それにより，日本語非母語話者であっても，日本語で，論説文や意見文のように長文ではないものの，自分自身が経験した，あるいは自分が見聞きした起承転結のある状況を，ある程度の長さの文を使い，他者に正確に順序立てて伝えなければならないことが多くなった。それに伴い，日本語教育においても，日本語でこのようなストーリー描写をする練習が必要であり，そして，各学習者の到達度を測る評価基準が必要であると考えられる。そこで，本研究では，「ストーリーを知らない読み手に対して，理解に不可欠な情報が不足することなく，前後関係などが正確に伝達されるよう書いて説明すること」を「ストーリーライティング(story writing；以下，SW)」と定義し，日本語教育現場で応用することを目的に，SW の評価基準の開発を試みた。なお，本研究で扱うのは，書き手が創造的に文章を書くクリエイティブライティング(creative writing)とは異なるもので，書き手の意見や感想は含まないものである。

SW に限らず，学習者の「書く」能力の測定には，実際の言語使用のパフォーマンスを評価できる点で高い真正性が確保できるというメリットがある。一方，評価に評価者の主観が入りやすいというデメリットもある(Backman，1990)。そのため，このような評価では，評価基準を作成し，評価者がそれに従って評価を行うことで評価者の主観をできるだけ排除した上で，一定の信頼性が確保できる評価基準であるかを確認する作業が行われることが一般的である。そこで，本研究でも評価の観点や到達段階を明示した評価基準を開発し，その基準をもとに評価の試行を行い，基準の有用性を検討することにした。

2. SW の評価基準開発の経緯

本研究で扱う SW の特徴として，まず読み手からの即時的な反応が得られにくいことが挙げられる。SW では対面で会話をしている場合のようにコミュニケーションの成否を確認する情報の受け手からのフィードバックが起こりにくいと考えられる。そのため，ストーリーの順序が仮に事実と異なって伝達されても，それを修正する機会が少なく，誤解が生じやすい

と考えられる。また，書き言葉特有の特徴として，読み手に伝えるべき情報を全て言語化しなければならない(Tannen, 1982)点もある。SW の場合で考えると，伝達されるストーリーが起こった場面に読み手が立ち会っていないことが普通であるため，伝えるべきストーリーの前提となる情報が伝わりにくいことから，書き手は読み手が状況を想像しやすいように使用する語や表現を工夫する必要がある。一方，意見文で重視されるであろう，読み手に対して説得的に論が展開できたか，書き手の主張が一貫していたか，内容が豊かで興味深いものであったか，といった観点は SW では重視されないポイントとなる。このように，「書く」ことの評価といっても，何を書くのかという課題の種類によって，評価のポイントとなる部分は異なると考えられる。

日本語教育において，学習者の「書く」能力の測定やそれについての研究はこれまでも行われてきている。菊池(1987)では，初級後半から中級レベルの日本語非母語話者向けの作文評価項目として，「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 項目があることを提案している。また，田中・坪根・初鹿野(1998)では，日本語教育に関わっていない日本語母語話者と日本語教師を対象に，日本語学習者の作文(意見文)を評価するポイントを調査した。その結果，学習者の作文の評価で焦点があてられる因子として，「正確さ」，「形式・構成」，「内容」，「豊かさ」が抽出されたと報告されている。これ以前の研究では，学習者の日本語の作文評価項目が詳細に示されたことはほとんどなく，菊池(1987)，田中・坪根・初鹿野(1998)は日本語学習者の「書く」能力を測定する際の評価項目を考えるにあたって，参考になるものであるといえる。しかし，これらの研究では，提示された評価項目に対して，到達度やレベルを判定する具体的な記述が提示されているわけではない。

一方，学習者にとって一種の目標となる「よい作文」とはどのような条件を備えている作文か，good writing の観点から日本語学習者の作文を研究している研究もある(例えば，田中・坪根，2011)。この種の研究では，複数名の日本語学習者の作文を評価の高い順に順位づけし，高評価群と低評価群を比較し両者の相違点を調べたり，もしくは高評価群の特徴を調べたりする方法で「よい作文」の特徴を質的に明らかにしようとしているものが多い。しかしながら，この高評価が得られた「よい作文」は他の学習者の作文を含め相対評価した結果としての「よい作文」である。したがって，評価は各研究での評価対象となった作文の質に依存する可能性があり，これらの研究結果で明らかになったことのみから学習者の作文の評価基準を作成することはできない。

以上のように，日本語で「書く」課題の種類は多様であることを考えると，それらを網羅するような汎用的な評価項目・評価基準を作成することは難しく，ある程度，課題の種類を限定してそれに特化した評価基準を作成せざるを得ない。本研究で扱う SW は，近年，日本語教育研究でアカデミック・ライティングの評価に関わる研究がさかんに行われている中であまり研究の対象とされてこなかった種類の課題であり，詳細で運用可能な評価項目・評価基準も作成されているとはいえない。そこで，以上の「書く」能力測定に関わる先行研究を参考に，SW の評価項目・評価基準を作成することにした。

3. SW 評価基準の作成

3.1 SW に類似したタスクの評価基準の概観

日本語学習者の「書く」能力の測定に使用するタスクとしては，小論文，意見文が扱われることが多い。本研究で扱うような SW 課題はほとんどなく，管見では，日本語教育分野で 1 件，心理学分野で 1 件のみであった。そこで，本研究で扱う SW に近い課題で「書く」能力を測定している例とそこで示されている評価項目を参照することにした。

まず，日本語教育分野で SW に近いタスクとして，『日本語教育のためのタスク別書き言葉コーパス』(金澤, 2014)に収められているタスク 8 がある。このタスクでは，友人とのケータイメールのやりとりの中で，友人に対して「鈴木先輩」に起こった大変な出来事を伝えるという，SW に比較的近いプロンプト(課題文・指示文)が受験者に与えられている。タスクでは，友人に伝えるべき先輩に起こった出来事が 4 コママンガで示されており，それを受験者が言語化して友人に伝える必要がある。このタスクの評価項目は，(同コーパスの他のタスク同様)「タスクの達成」，「タスクの詳細さ・正確さ」，「読み手配慮」，「体裁・文体」と「総合評価」で，この他に評価のポイントが示されていた。それぞれの項目について，評価のポイントをもとに○，△，×の 3 段階，または○，×の 2 段階で評価をすることになっていた。金澤(2014)の評価項目は，同じコーパスに収められている他のタスクにも共通して用いられる項目であったということもあり，SW の構成要素を強く反映しているものではない。また，このタスクでは，4 コママンガの内容を友人という親しい人物に対して報告するため，親しい友人に対して，親しみが感じられる表現を用いているかが 1 つのポイントとなっている。さらに，ケータイの返信メールであるという設定のため，適切な書き出しの文が使用されているかということも「体裁・文体」のところで評価されることになっている。なお，金澤(2014)では，以上

のような評価基準で 3 名の評価者が評価を行い，意見が分かれた部分については話し合いを通して意見が一致したところを最終評価としている。

次に，心理学分野で SW に近いタスクとして，『標準失語症検査(改版 第 2 版)』(日本高次脳機能障害学会，2003)で使用される「まんがの説明」というタスクがある。これは，日本語で行われるタスクであるが，基本的には日本語母語話者に対して実施するタスクである。「まんがの説明」のタスクは本研究で実施する SW と同じく 5 コマからなるマンガのストーリーを描写するというもので，独立した「語彙」「文法的誤り，文字の誤り」「なめらかさ」の 3 つの評価項目をそれぞれ 6 段階で評価する。3 つの項目の各段階には，例えば「文法的誤り，文字の誤り」の項目では，「偏・つくり，濁・拗・促音，送り仮名の誤りが 1 か所」であれば 6，「偏・つくり，濁・拗・促音，助詞・語尾，字性錯書，送り仮名の誤りが 2 か所以内」であれば 5 というように具体的な説明があり，そのいずれにも該当しない場合を 1 としていた。ただ，評価項目「語彙」の場合には，このタスクでマンガを説明するために必ず使用すると考えられる基本語彙が設定されており，その基本語彙あるいは意味が類似している関連語彙がいくつ産出されたかが評価の段階を判断するポイントとなっているため，この評価基準を他のマンガの状況描写の課題に応用するには，別途，各課題に合った基本語彙を提示する必要がある。

以上の 2 つのタスクの評価基準を参考に，本研究では日本語の SW による「書く」能力を汎用的に評価する基準を作成する。

3.2 評価方法の選定

パフォーマンス評価には，大きく分けて総合的評価(包括的評定; holistic rating)と分析的評価(分析的評定; analytic rating)と呼ばれる評価方法がある。前者が，パフォーマンスの印象を全体的なものとして記録する方法であるのに対し，後者はそのパフォーマンスのさまざまな面について個別にアセスメントを行う方法である(McNamara, 2000)。両者は，それぞれ評価の目的によってメリット・デメリットがあると考えられるが，本研究では SW の評価基準を日本語教育における SW の教育に応用されることを目的として開発していること，また，日本語で SW ができるようになることに日本語のどのような能力が貢献しているのかを知ることが第二言語としての日本語の習得研究を行う上で重要であると考えることから，SW を多角的に評価する分析的評価を行うことにした。

分析的評価では，パフォーマンスのさまざまな面について，各教育機関やコースの目的

に応じて，合計で 100%になるようにそれぞれ重みづけをして評価されることが多い(田中, 2016)が，本研究では，その中でも SW に含まれると考えられる各特性について重みづけをすることなく独立したものとして評価していくマルチプル特性評価(multiple trait scoring)[5]で SW を評価することにした。

3.3 評価の特性の設定

マルチプル特性評価で学習者のパフォーマンス(ここでは SW)を評価しようとする場合，評価の特性(trait)と呼ばれる，評価の観点を設定する必要がある。評価の特性をたてるにあたっては，日本語で SW ができることには，どのような能力が関わっているか，いわゆる SW の構成要素を考えなければならない。先行研究では，L2 のライティング能力は，文章の構成や内容に関わる能力(writing expertise)と L2 の言語能力(L2 proficiency)で構成されている(Cumming, 1989)が，本研究で扱うような SW の場合は，具体的にどのような能力で構成されていると考えられるのだろうか。英語教育分野では，Chiang(2003)が作文評価における評価特性を「談話的特性(discourse features)」と「文法的特性(grammatical features)」に分類し提示した。本研究でも，この枠組みにならい，SW の特性を 4 項目挙げる。

まず，Chiang(2003)の「談話的特性」に分類される特性として，本研究では「伝達性」と「結束性」の 2 項目を挙げる。「伝達性」とは，伝えるべき情報が過不足なく伝わっているかというものである。SW では，ストーリーの展開を実際の順序に忠実に表現し，状況に立ち会っていない(課題の絵を見ていない)読み手に対して，状況を正しく伝えることが重要となる。また，それだけではなく，読み手に正しく状況を伝えるためには，ストーリーの前提となる情報も補いながらストーリーの展開を説明していく必要がある。これらが十分にできているかを「伝達性」とし，1 つの特性とした。一方，「結束性」とは，文と文とに結束性があり，文章にまとまりがあるかというものである。SW では，1 つの状況を説明するにあたって文がねじれていないだけではなく，一連のストーリーの複数の場面のつながりを接続詞などを効果的に使用したり，視点を統一させたりして，ストーリー内の動作主や因果関係を誤解なく伝える必要がある。これらが十分にできているかという観点を「結束性」とし，1 つの特性とした。

次に，Chiang(2003)の「文法的特性」に分類される特性として，本研究では「表現の豊か

[5] 本研究では，マルチプル特性評価を分析的評価の一種であると考える Weigle(2002)の定義に従っているが，マルチプル特性評価は総合的評価に対するものとして捉えている研究(Hamp-Lyons, 1991)もある。

さ」と「表現の正確さ」の 2 項目を挙げる。「表現の豊かさ」とは，SW の読み手が状況を理解しやすいように，場面に合った多彩な表現が選択されているかということである。日本語での SW であれば，オノマトペや複合動詞を使用するなどの方法で，伝えたい状況をよりわかりやすく伝えることができると考えられる。このような点ができているかという観点を「表現の豊かさ」とし，1 つの特性とした。一方，「表現の正確さ」とは，SW で使用される表現が語彙的・文法的に正確であるかということである。SW を構成する要素として，SW で使用しているコロケーションが正しいか，動詞・形容詞などの活用は正しいか，またコンピューターで入力する場合，表記のミスはないかなどが，表現の正確さに関連すると考えられる。このような点ができているかという観点を「表現の正確さ」とし，1 つの特性とした。

以上のように，本研究では，日本語による SW を「伝達性」，「結束性」，「表現の豊かさ」，「表現の正確さ」の 4 つの評価のための特性で評価することにした。

3.4 評定尺度の設定

上述の 4 つの SW 評価のための各特性について，到達度のレベルを示す評定尺度を設定した。「書く」能力評価の評定尺度はさまざまであり，評価項目によってその尺度の数が異なる場合もあるが，通常 3〜9 のレベルが設けられることが多い(McNamara, 2000)。本研究の SW 評価では，田中・坪根・初鹿野(1998)ほか一連の研究で用いられているアカデミック・ライティングの評定尺度，そして標準失語症検査で使用される「まんがの説明」のタスクの評定尺度を参考に，6 レベルの評定尺度を設定することにした。

評定尺度の設定にあたっては，各特性の評定尺度を詳細に記述するのに先立って，各レベルのおおまかな到達度を設定し，特性間のレベルの整合性を取ることを目指した。本研究の SW 評価基準の各レベルのおおまかな到達度は，表 1 に示した通りである。評定尺度のレベル 6 をネイティブレベルとし，レベル 6 からレベル 2 までのいずれにも該当しないものをレベル 1 とした。次に，レベル 5 とレベル 4 を非母語話者に慣れていない読み手であっても理解できるレベルであると定義し，レベル 3 とレベル 2 を非母語話者に慣れている読み手(例えば，言語教師など)であれば理解できるレベルであると定義した。さらに，各特性の不足(例えば，伝達性の不足)により誤解が生じうるか，読み手にとってストーリーの内容理解にどの程度負担がかかるかという観点から各レベルの到達度を設定した。なお，先行研究(田中・坪根・初鹿野，1998; 金澤，2014)で評価項目または特性として挙げられていた「読み手への配慮」に関しては，SW の課題の特質上，特定の読み手へ向けたものか不

特定の読み手へ向けたものかを含め具体的に読み手を想定することが困難であったため，特性の１つとして設定することはしなかった。しかし，SW においても読み手にとって内容理解がしやすい文章が書けたかという点は重要な点であると考え，SW の到達度を記述する際に，この点についても考慮に入れ記述することにした。

表1　SWにおける各レベルのおおまかな到達度

レベル	おおまかな到達度
6	ネイティブレベル。読み手を意識して理解を容易にする工夫を多用している。
5	非母語話者に慣れていない読み手でも理解可能である。 各特性の不足もわずかである。
4	非母語話者に慣れていない読み手でも理解可能である。 各特性の不足により，読み手に誤解や理解への負担を生じさせる場合もある。
3	非母語話者に慣れている読み手であれば理解可能である。 各特性の不足により，読み手に対して相応に負担がかかったり， 内容理解において誤解を生じさせる可能性がある。
2	非母語話者に慣れている読み手であれば理解可能である。 特性の不足により，読み手に対して大きな負担がかかったり， 内容理解において誤解を生じさせる可能性が高い。
1	上記のいずれにも該当しない。

表１に示した各レベルのおおまかな到達度に基づいて，各特性の到達度を記述した(補記１参照)。その際に，本基準を SW の汎用性のある評価基準にするために，特定のストーリーの状況描写にのみ対応する表現を記述に含めることがないように配慮した。

4. SW 評価の試行と有用性の検討

4.1 評価対象

本研究では，作成した SW 評価基準の信頼性を確認するために，作成した評価基準に基づいて評価を試行した。評価に用いた作文は，国立国語研究所によって収集され一般公開されている「多言語母語の日本語学習者横断コーパス(I-JAS; International Corpus of Japanese as a Second Language)」(http://lsaj.ninjal.ac.jp/?cat=3)に収録されている SW の課題である(補記 2 参照)。このコーパスのデータ収集に協力した日本語学習者は，日本語能力

を測るためのテスト結果と日本語での 7 種類(そのうち 2 種は任意)の話し言葉・書き言葉のデータを提供している。本研究で使用した SW 課題は 2 タスクあり，その両方を評価対象とした。タスク 1 は 5 コマからなる「ピクニック」で，タスク 2 は 4 コマからなる「鍵」である。それぞれのタスクの登場人物の名前(「ケン」「マリ」)，難易度が高くストーリー描写には不可欠な語とその英訳(「バスケット(basket)」「警官(policeman)」など)，ストーリー描写を行う上で最初の 1 文にあたる文(タスク 1 では「朝、ケンとマリはサンドイッチを作りました。」，タスク 2 では「ケンはうちの鍵を持っていませんでした。」)はあらかじめ提示されている。迫田(2016)によると，この課題は，各学習者がパソコンに入力する形で行われ，1 タスクにつき 10 分で解答するように指示された。なお，辞書の使用は許可されていない。本研究では，前述のコーパスに第四次公開版として公開されている，日本語学習者 650 名および日本語母語話者 50 名の計 700 名の作文を 2 タスク，計 1,400 本の SW の作文を評価した。作文を書いた日本語学習者の母語は，インドネシア語，スペイン語，タイ語，トルコ語，ドイツ語，ハンガリー語，フランス語，ベトナム語，ロシア語，韓国語，中国語であった。ただし，英語については，母語が中国語である日本語学習者が多かったので，母語を英語と特定することができなかったので，含まなかった。

4.2 評価者と評価方法

評価の試行にあたっては，2 名の評価者(Y, S)がそれぞれ 1,400 本の作文を全て評価した。同評価者は，日本語教育経験 10 年以上の日本語教師で日本語母語話者である。評価は，先に挙げた評価基準を各自熟読し，「伝達性」，「結束性」，「表現の豊かさ」，「表現の正確さ」の 4 つの特性について，それぞれレベル 6 であれば 6 点，レベル 5 であれば 5 点というように得点をつけていく，マルチプル特性評価(分析的評価)を行った。

4.3 SW 評価基準の有用性の検討

4.3.1 評価結果と評価者間信頼性

本研究で扱う SW のようなパフォーマンス評価では，評価に評価者の主観が入りやすいこと (McNamara, 2000)により，評価者による評価が信頼できるかが問題になることが指摘されている。そこで，まず，本研究の評価者 2 名の評価結果を概観し，両者の評価の傾向をみてみる。表 2 に示すのは，評価者 Y と評価者 S の 2 種類の SW の評価の平均をまとめたものである。

表2を参照すると，SW1では評価者Yの合計点平均と評価者Sの合計点平均には，0.81点の差がある(S>Y)ことがわかる。同様に，SW2 でも評価者 Y の合計点平均と評価者 S の合計点平均には，1.04 点の差がある(S>Y)ことがわかる。つまり，評価者 Y は評価者 S より全体的に厳しい評価をしていることになる。特に，SW1 でも SW2 でも「正確さ」という特性の平均点の差がやや大きく(評価者 Y の方が評価者 S よりややばらつきがある評価をしており)，この部分の差が両者の合計点平均の差につながっていると考えられる。このように，2名の評価者の評価が完全に一致することは難しいが，両者の差がどの程度，評価者間の信頼性に影響を与えるものか，より詳しくみていく必要がある。

表2　2名の評価者によるSWの評価結果一覧

	評価者	*M* と *SD*	伝達性	結束性	豊かさ	正確さ	合計
SW1	Y	*M*	2.72	2.65	2.77	2.59	10.72
		SD	1.02	0.98	1.08	1.19	3.98
	S	*M*	2.92	2.81	2.89	2.90	11.53
		SD	1.05	1.02	1.01	1.09	3.88
SW2	Y	*M*	2.72	2.68	2.76	2.52	10.69
		SD	1.07	0.95	1.09	1.15	3.98
	S	*M*	2.99	2.90	2.92	2.91	11.73
		SD	1.10	1.11	1.06	1.10	4.11

注: *N*=700. *M* は平均，*SD* は標準偏差.

表 3 は，評価者 Y と評価者 S それぞれの評点について，合計得点と 4 つの特性ごとの得点についてのピアソンの積率相関係数を計算した結果をまとめたものである。

表3　2名の評価者間の相関

	SW1		SW2	
特性	相関係数	*p*	相関係数	*p*
伝達性	0.72	***	0.76	***
結束性	0.67	***	0.69	***
豊かさ	0.79	***	0.80	***
正確さ	0.80	***	0.79	***
合　計	0.82	***	0.84	***

注: *N*=700.

評価者 Y と評価者 S の合計得点の相関係数は，SW1 で 0.82(p<.001)，SW2 で 0.84(p<.001)であった。評価者間一致度で許容できる範囲は，相関係数が 0.7 から 0.9 である(McNamara, 2000)ことを考えると，この SW 課題における評価者間信頼性は確保されているといえる。また，特性別の相関係数をみても，「伝達性」「(表現の)豊かさ」「(表現の)正確さ」については許容範囲であると判断できる相関係数であり，特性別にみても，その評価者間信頼性は，確保されているといえるだろう。しかしながら，「結束性」については，わずかに 0.7 を下回っており，許容できる評価者間一致度におよばなかった。

4.3.2 日本語能力測定としての SW 評価基準の妥当性の検討

L2 のライティング能力は，文章の構成や内容に関わる能力(writing expertise)と L2 の言語能力(L2 proficiency)で構成されている(Cumming, 1989)ため，必ずしも日本語の能力のみが日本語の SW の評価に反映されるとは限らない。しかしながら，日本語教育でこの SW を扱おうとする以上，日本語能力をある程度反映している SW の評価基準であることが望まれる。このような，測ろうとするものが測れるようにテストが作られているかどうかの程度を表す概念を妥当性という。従来は，妥当性は内容的妥当性，構成的妥当性，基準連関妥当性に分けて捉えられてきた。しかし，現在ではこれらを単一の概念として捉え，構成概念妥当性を中心に据えて，それを確認する方法がとられており，新しいテストの妥当性を確認するために同様の構成概念を有する既存のテストと関連づけて検討されることがある(野口・大隅, 2012)。これを踏まえ，本研究では，SW での評価が既存の日本語能力測定テストと高い相関があるかを確認することで，評価基準の妥当性を検討した。

本研究で，SW 評価の試行に用いた「多言語母語の日本語学習者横断コーパス」に収録されているデータ提供者である学習者は，SW の他にも日本語能力測定のテストを受験し，その結果を提供している。彼らの，J-CAT と SPOT の得点は表 4 の通りである。

表4　日本語学習者の日本語能力テスト結果

日本語能力テスト	平均	標準偏差
J-CAT聴解(100点満点)	52.69	16.60
J-CAT語彙(100点満点)	53.25	16.94
J-CAT文法(100点満点)	47.69	16.42
J-CAT読解(100点満点)	45.91	12.98
J-CAT合計(400点満点)	199.54	53.47
SPOT得点(90点満点)	66.43	10.91

注: N =650.

次に，表4に示した日本語能力テストの結果とSWの評価とのピアソンの積率相関係数をまとめたものを，表 5 に示す。

表5 2名の評価者によるSW評価結果と日本語能力テストの相関

	YSW1合計		YSW2合計		SSW1合計		SSW2合計	
日本語能力テスト	相関係数	*p*	相関係数	*p*	相関係数	*p*	相関係数	*p*
J-CAT聴解(100点満点)	0.50	***	0.50	***	0.53	***	0.54	***
J-CAT語彙(100点満点)	0.41	***	0.46	***	0.45	***	0.46	***
J-CAT文法(100点満点)	0.40	***	0.45	***	0.43	***	0.42	***
J-CAT読解(100点満点)	0.43	***	0.43	***	0.47	***	0.43	***
J-CAT合計(400点満点)	0.51	***	0.54	***	0.55	***	0.55	***
SPOT得点(90点満点)	0.45	***	0.51	***	0.51	***	0.50	***

注: N=650.

表 5 によると，評価者 Y と評価者 S の各タスク(SW1 と SW2)の評価結果と各日本語能力測定テストの間には，いずれも 0.4 から 0.6 の間のある程度高い相関があることがわかる。このことから，SW ができるということは必ずしも日本語能力のみを反映しているものではないが，一般に SW で高評価を得る学習者の日本語能力テストの結果は高得点であるといえそうである。したがって，SW で評価されている特性やその評価方法は，概ね妥当であると考えられる。ただし，J-CAT および SPOT が測定する日本語の知識・技能はそれぞれ異なるが，日本語能力の中でも特にどのような知識・技能と SW 評価の結果に強い関係があるのかは，この分析結果のみで結論づけることはできない。この点は今後の課題としたい。

4.3.3 SW 評価におけるタスクの影響

最後に，SW 評価に SW のタスクの影響がみられるかを検討した。パフォーマンス評価の不一致に影響するといわれているのは，スクリプト，評価基準や評価方法，評価者などがある(田中，2016)。その他にも，日本留学試験の「記述問題」におけるトピックの影響について調べた廣瀬(2008)では，受験者に与えられたトピックによって日本語のライティング能力評価に違いがみられることを指摘している。本研究で評価に使用した SW のタスクには 2 種類あるが，タスクによって SW 評価が変わることはあるのだろうか。表 6 は，2 名の評価者による SW のタスク間相関係数をまとめたものである。

表6　2名の評価者によるSWのタスク間相関

	YSW1合計	YSW2合計	SSW1合計	SSW2合計
YSW1合計				
YSW2合計	0.87 **			
SSW1合計	0.82 **	0.79 **		
SSW2合計	0.74 **	0.84 **	0.77 **	

注: N =700.

表 6 をみると，評価者 Y による評価(YSW1 合計と YSW2 合計)では相関係数が 0.87 (p<.01)，評価者 S による評価(SSW1 合計と SSW2 合計)では相関係数が 0.77(p<.01)と，どちらの評価者の場合もタスク間相関は高く，タスクによって評価が大きく異なることはないことが明らかになった。この結果は，廣瀬(2008)の結果と異なるが，それは SW という課題の特徴にも関連している可能性がある。本研究で扱っている SW では，書き手の立場の一貫性や意見・趣旨の明確さは評価の対象とならず，あくまで事実関係を正確に描写することが求められる。そのため，日本語能力以外の内容に関わる能力の影響を受けにくく，評価者間およびタスク間の相関が高くなったのではないかと考えられる。

5. おわりに

本研究では，近年，日本語非母語話者にとっても重要性が高まったSWの到達度を評価し，日本語教育における SW の指導に役立てる目的で，SW の評価基準の開発を行った。評価の試行を行い，その有用性を検討したところ，信頼性・妥当性両面から本研究で開発した評価基準はある程度有用なものであると判断できる結果が得られた。一方で，例えば SW の「結束性」の評価の一致度は満足がいく高さであるとはいえず，評価基準の改善の余地もあることがわかった。また，本研究では評価の過程での評価の不一致が起こりやすい面について，質的に分析することはできなかった。今後，本研究における SW の評価基準をより有用なものにするには，これらの点についても再考していく必要があると考えられる。

［参考文献］

Backman, L. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press. (バックマン, ライル F., 池田央・大友賢二 (監修)大友賢二・笠島準一・服部千秋・法月健 (訳) (1997)『言語テストの基礎』東京：C.S.L.学習評価研究所.)

Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning, *System, 31*, 471-484.

Cumming, A. (1989). Writing expertise and second-language proficiency, *Language Learning, 39(1)*, 81-141.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp.241-276). Norwood, NJ: Ablex.

McNamara, T. (2000). *Language testing*. Oxford, UK: Oxford University Press. (マクナマラ, ティム, 伊東祐郎・三枝令子・島田めぐみ・野口裕之 (監訳) (2004)『言語テスティング概論』東京：スリーエーネットワーク.)

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives, *Language, 58(1),* 1-21.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

金澤裕之(編)(2014)『日本語教育のためのタスク別書き言葉コーパス』東京：ひつじ書房.

菊池康人 (1987)「作文の評価方法についての一私案」『日本語教育』63, 87-104.

迫田久美子（2016）『海外連携による日本語学習者コーパスの構築―研究と構築の有機的な繋がりに基づいて―I-JAS 構築に関する最終報告書』（平成 24～27 年度科学研究費助成事業（基盤研究 A）研究成果報告書）

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』(pp.34-53), 東京：くろしお出版.

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.

日本高次脳機能障害学会 Brain Function Test 委員会 (2003)『標準失語症検査マニュアル(改訂第 2 版)』東京：新興医学出版社.

野口裕之・大隅敦子 (2012)「テスティング・評価」近藤安月子・小森和子(編)『研究社日本語教育事典』(pp.337-360), 東京:研究社.

廣瀬香恵 (2008)「日本留学試験『記述問題』におけるトピックの影響」『日本語教育』136, 59-67.

大和 祐子 - 大阪大学 日本語日本文化教育センター・准教授

玉岡 賀津雄 – 名古屋大学大学院 人文学研究科・教授

斉藤 信浩 - 九州大学 留学生センター・准教授

【補記 1：SW の各特性(トレイト)のレベル別記述】

伝達性	
6	・絵を見ていない読み手でも，問題なく状況や場面が想起できる。 ・必要に応じて絵に含まれている情報以外の情報を入れつつ，適切にストーリーを描写している。 ・読み手が内容を理解しやすい構成で，かつ効果的な表現が多用されている。
5	・絵を見ていない読み手にも，ストーリーの大部分の状況や場面が誤解なく想起できる。 ・必要に応じて絵に含まれている情報以外の情報を入れているものの，それには過不足がある場合がある。 ・読み手の内容理解に配慮が感じられる構成で，効果的な表現も使用されている。
4	・ストーリーの詳細は読み手にとって理解できない部分があるものの，ストーリーのおおまかな流れは十分に理解できる。 ・表現力の不足，正確さの不足などにより，誤解を招く場合がある。
3	・非母語話者が書く文を読み慣れている人であれば，ストーリーの状況が理解可能であるが，読み手の内容理解にかかる負担は大きい。
2	・非母語話者が書く文を読み慣れている人であれば，各文の意味とストーリーのおおまかな流れがなんとか理解できる。ただし，読み手の想像力に頼るところが大きい。
1	以上に該当しない。

結束性	
6	・場面と場面のつながりが接続表現の使用などによって明確に表現できている。 ・書き手の視点が統一され，読み手に誤解なく状況が伝達できている。
5	・場面と場面のつながりは接続表現の使用などによってほとんどの部分で問題なく表現できている。 ・書き手の視点が統一されていない部分があるが，意味の理解には支障がない。
4	・実際のストーリー展開と書かれた文章の展開とに若干の齟齬があり，誤解を招く恐れがある部分がある。 ・各場面の情報は複文も用いて表現されているが，場面と場面のつながりを正確かつ十分に表現できていないところがある。 ・書き手の視点が統一されていない部分があるが，意味の理解は可能である。
3	・接続表現や授受表現が適切に使われていないことにより，書かれている文章の前後関係とストーリーの前後関係が異なる箇所がある。 ・書き手の視点が統一されていないために，読み手に負担がかかる部分がある。
2	・文レベルでは，一部文のねじれなどは見られるものの，概ね理解可能な文を書いている。
1	以上に該当しない。

表現の豊かさ	
6	・オノマトペや副詞が場面状況を説明する上で効果的に使用されている。 ・伝聞や引用が効果的に使用され，的確に状況が表現されている。 ・それぞれの絵に対してストーリーを説明する上で過不足のない情報が的確な語で説明されている。
5	・オノマトペや副詞が場面状況を説明するために使用されているが，若干の不自然さがある。 ・伝聞や引用が使用され，状況が理解しやすくなる工夫が見られる。 ・それぞれの絵に対してストーリーを説明するために多様な語を用いているが，情報に過不足がある部分がある。
4	・場面状況を説明するための語が不足している部分があるが，意味理解にはほとんど問題がない。 ・伝聞や引用が一部正確ではない部分があり，読み手によっては誤解を招く可能性がある。
3	・場面状況を説明するための語が不足しており，読み手に不正確な場面を想起させる可能性がある。
2	・使用する語などが適切ではないことにより，読み手には場面理解が困難になる部分がある。 ・状況を描写するにあたって必要最低限の語を用いている。
1	以上に該当しない。

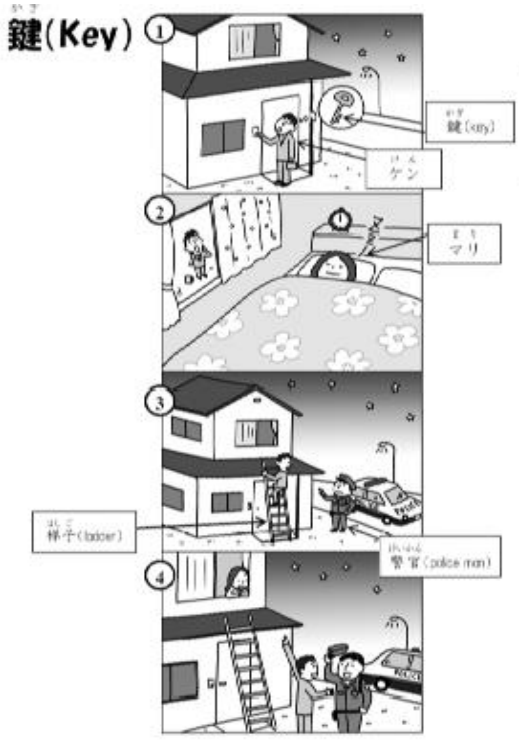
表現の正確さ	
6	・テンス／アスペクトの誤り，自動詞／他動詞の誤りがない。 ・助詞が適切に使用されている。 ・動詞（複合動詞を含む），形容詞などの語が意味的・文法的に正しく選択されている。 ・文体が統一されている。
5	・テンス／アスペクトの誤り，自動詞／他動詞の誤りはあるものの十分に読み手はストーリーの内容を理解できる。 ・大半の助詞は適切に使用されているが一部誤りがある。 ・意味的に大きな相違はない。場面状況に合わない語が一部選択されているが多くの語は正しく選択されている。 ・文体が概ね（１，２か所を除き）統一されている。
4	・テンス／アスペクトの誤り，自動詞／他動詞の誤りはあるものの，読み手はストーリーの内容を概ね問題なく理解できる。 ・ごく一部で活用の誤りが見られるが，意味の理解には影響はない。 ・意味的・文法的に適切ではない語の選択が目立つ。 ・文体は統一されていない部分がある。
3	・表記が不正確である語（長音，濁音の欠落など）が散見されるが，読み手の推測により意味の理解は可能である。 ・活用の誤りも複数あるが，意味は推測可能である。 ・文体が統一されていない。
2	・文法が不正確であることで，ストーリーの流れの理解に困難が伴う部分がある。 ・表記が不正確である語（長音，濁音の欠落など）が散見され，読み手が意味を誤解する恐れがある。 ・文体が統一されていない。
1	以上に該当しない。

【補記 2：I-JAS コーパスにおける SW 課題】　　迫田 (2016) pp. 63-64 より転載

絵をよく見てください。
そして、次のことばに続けて、この5つの絵のストーリーを話してください。
(story)

「ピクニック」

朝、ケンとマリはサンドイッチを作りました。



絵をよく見てください。
そして、次のことばに続けて、この4つの絵のストーリーを話してください。
(story)

「鍵」

ケンはうちの鍵を持っていませんでした。

Developing evaluation criteria for Japanese narrative abilities using the corpus data of the International Corpus Japanese as a Second Language

YAMATO, Yuko
(Associate professor, Center for Japanese Language and Culture, Osaka University, Japan)
TAMAOKA, Katsuo
(Professor, Graduate School of Languages and Cultures, Nagoya University, Japan)
SAITO, Nobuhiro
(Associate professor, International Student Center, Kyushu University, Japan)

Abstract: In this study, we report a series of procedures for developing evaluation criteria for narrative abilities for the applicable use of Japanese language education. First, based on previous studies, we established the four categorical characteristics of narrative abilities in Japanese: (1) transferability, (2) cohesion, (3) richness of expression, and (4) accuracy of expression. Then, for each characteristic, we developed criteria consisting of six levels (six points). In order to examine the usefulness of the evaluation criteria, we evaluated texts of the story writing (SW) tasks contained in the corpus data of the International Corpus Japanese as a Second Language (I-JAS). The result showed that the correlation (inter-evaluator reliability) between the two evaluators was very high as r= 0.82 for the SW1 and r=0.84 for the SW2 (see Table 3). The evaluation scores of SWs showed relatively high correlations with Japanese language proficiency tests ranging from 0.41 to 0.55 (see Table 5). Furthermore, the evaluation results for the two SWs indicated high correlations from 0.77 to 0.87 (see Table 6). From these results, the evaluation criteria of narrative abilities developed in this study is applicable for the use of Japanese language education.

Keywords: story writing (SW), evaluation criteria, reliability, International Corpus Japanese as a Second Language (I-JAS), narrative ability