英文エッセイの自動レベル判定システムと手動採点結果の比較検証：CEFR-Jライティング・テストタスク構築ための予備調査

中　谷　安　男

1．はじめに

本論は根岸科研におけるCEFR-Jライティング・テストタスク開発プロジェクトの予備調査結果の一部を報告するものである[1)]。ここでは，３つのレベルのタスクを実施した結果を基に，英語エッセイの自動レベル判定システムと手動採点結果を比較する。適合性や，手動採点方法の改善，さらに機械学習による自動システム向上への示唆を行う。

2021年から大学の入学選抜方式の改革が本格的になり，これまでのセンター試験の代わりに，英語の４技能を総合的に評価できる試験を導入することになった。大きな変更点は，主体的に表現する力として「話す」，「書く」能力が試される。

確かにグローバル社会の発展に伴い，英語のライティング技能は日本人にとって重要性は増すであろう（中谷, 2010; Nakatani, 2016）。国際社会において，互いに理解し合い，合意をした証として文書を交わす。特に，国際ビジネスにおいては英語が標準語で，交渉の前に電子メールなどで連絡を取る。また口頭で合意を得ても，最後は英文の契約書を結ぶことになる。このように，英語のライティング力は国際交流が進む社会で活躍する人材にとって重要な素養となる（Nakatani, 2015; 中谷，2016）。

しかし，ライティングテストの入試制度への本格的導入に際して，採点業務や評価方法の複雑さやコストの問題がある。例えば2019年度 センター入試利用者は57万6,829人となり，英語科目の受験者は53万7,663人である[2)]。これだけ大量の受験者が英文エッセイ・ライティング試験を仮に全員受けると，その採点業務や結果産出の適切な実行には多くの課題がある。全てを手動で採点するのは実質的に不可能である。いずれ受験生はCEFR（Common European Framework of Reference for Languages）に準拠した，いくつかの民間テストの結果を活用することになる予定である。これらの業者は，コンピュータによる自動レベル判定システムを当然のように活用することになる。

自動レベル判定の開発や導入に関しては，近年様々な取り組みが行われている。代表的なシステムとして米国の ETS（Educational Testing Service）が開発した *e-rater* あり，TOEFL iBT や GMAT のライティングテスト結果判定に活用されている（Enright and Quinlan, 2010）。また英国などの大学留学に活用される IELTS のトレーニングにも活用できる *Cambridge Assessment English* が開発した *Write & Improve* がある。これは，レベルによって表示されたタスクを選択し，英文を入力すると CEFR レベルの判定が得られる[3)]。

主にこれらのシステムは，被験者が書いた英文の単語数や文数，語彙の難易度，多様さ，構文の複雑さなどを統計的に分析して判定を行う（例 Crusan, 2010; Lu, 2017）。しかし汎用性のあるシステム構築の過程で，様々な母語を持つ被験者の英文ライティング採点データを活用している。つまり，日本の英語学習者に焦点を当て構築されたものではない。場合によっては，その国特有の言語環境に適応する英文エッセイの採点や，エラー分析を活用した自動レベル判定システムの方が，より精密な予測が可能となると思われる。

自動システムが改善され，様々な検証が行われているが，ライティング評価を全て自動で行うことに対する妥当性や信頼性に関しては議論がある

（例　水本，2008）。現状では，まだ人手による採点と機械による自動レベル判定の融合が望ましいと言える（Crossley1，Roscoe，and McNamara，2013）。手動採点の実施課題としては，採点者のトレーニングの問題がある。大量のテストデータを効率よく採点し，適切な判定を下せるには，かなりの訓練が必要となる。しかしながら，これまで日本人を対象とした英語ライティングテストの採点者トレーニングに関する報告はあまり多くない（例　根岸，2012）。

以上の観点から，本論は日本の英語学習者向けに構築されたCEFR-Jにおける，ライティング・テストタスクの採点基準や採点方法，さらにタスクのレベルの適性を確認していく。この際，手動による採点結果と，CEFR-J用に構築された自動レベル判定システムによる結果の比較を通して考察を行う。

２．研究の背景

2.1　CEFRとCEFR-J

CEFRとは，言語の学習・教育・能力評価の基本的枠組みであり，欧州評議会加盟国共通の外国語習熟度の尺度として広く活用されている（Nakatani，2017a，2017b）。歴史的には，1971年より多言語国家の集合体であるヨーロッパ連合（EU）において，各国市民の言語能力の向上と，連合内におけるグローバルな評価基準の整備が開始された。加盟国内の教育関係者が理念を共有し，各言語に関する学習について共通の目標，内容，教授法を設定することが目的であった。例えば，スペインでの英語教育も，オランダでの英語教育も共通の尺度があれば，互いに汎用性があり，教育者同士の連携や協力が容易になる。1975年に，成人としてコミュニケーションに必要な，最低限の言語運用能力のための概念，及び機能シラバスである *Threshold Level* が発表された。次年度にフランス語版が作製され，英

語に関しては van Ek and Trim（1990）により体系化されたものが発表され，次第にCEFRの体系が構築されていった。

詳細な報告書 *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*（CEFR: LTS）（2001）に記載されているが CEFR の特徴は action-oriented である。つまり単に言語知識の習得ではなく，目標言語を使って何ができるかという点を強調している。言語運用能力の到達基準を，基礎段階の使用者（A: Basic User），自立した使用者（B: Independent User），熟達した使用者（C: Proficient User）と大まかに３段階に区別した。さらに，それぞれの到達基準を２段階のレベルに分け，A1からC2までの６レベルとした。この各レベルの内容は，Can-Do リストとよばれる「何ができるか」の指標で具体的に表わされている。以上のようなCEFRに基づいた到達度に共通認識を持つことで，目標言語能力の一定の証明になる（Nakatani，2013）。この結果，EU内での移住等の希望者に，受入の必要条件として客観的な言語の目標を示せ，学習に向けた公平な指標ともなる。

CEFR-J は，CEFR を日本の英語教育事情に適合させるために開発された（小池　他，2008）。日本においても，CEFR の指標導入の可能性について多くの研究が行われた（例　根岸，2008）。しかしアルファベット関連文字を主に使用し，言語の類似性もある欧州の学習者に適したものを，そのまま日本の学習に活用することは容易ではない。このため，小池科研において日本における英語教育の指標である CEFR-J の基本案が提示された[4]。続く投野科研の成果として，CEFR-J の具体的構成や Can-Do リストの初期版が完成された[5]。この大きな特徴として日本の英語学習者に適した，より初級レベルの Pre-A1 が加えられた。またA，Bレベルもよりも，細分化された。さらに続く投野科研では，学習者コーパスによる英語 CEFR レベル基準の特定が行われた[6]。この中で，Error Profile として，日本人学習者のJEFLL（Japanese English as a Foreign Language Learner）Corpus が活用された[7]。奥村・能登原両氏が中心となり，JEFLL Corpus の自動エ

ラータグ付与したデータを活用して，レベル判定に有効なエラー項目をSupport Vector Machine を用いて検出，その特徴の重み付けを評価した。これが本論で検証する自動採点システム構築の基礎となっている。

2.2　根岸科研

これまでの一連の CEFR-J 関連プロジェクトの成果に基づき，この科研では，Can Do descriptors に基づいたテストタスクの作成と検証を行っている。また同時に，英語のインプット・アウトプット両方のテキストの自動レベル判定などのツールを開発している。特に著者が担当しているのは，英語ライティングのテストタスクの作成と検証で，同志社大学の能登原祥之先生，玉川大学の工藤洋路先生と３名が中心になり取り組んでいる。これまで，PreA1からA1.1までのテストタスク作成と評価基準の構築，テスト実施，及び成果の検証を行った[8)]。

今回はこの３名によって，新たにCEFR-J のディスクリプタに合わせたA2.2.1からB1.2.2レベルの６つのテストタスクの試案を作成し，同時に採点方法も考案した。これを都内の私立の高等学校で実施してもらい，採点の結果を分析している過程である。本論はこれらの研究の予備検査として，今後の検証方法の妥当性を確認するために，結果の一部である３つのタスクに限り著者が分析したものを報告する。

2.3　ライティングの評価

英文エッセイなどは，書き手の属性や能力よって多様なライティングが産出される（Crusan, 2010）。このため，正確に評価を行うのはそれほど容易ではなく，一定のトレーニングなどが必要となる（Weigle, 2002）。特に，被験者が目標言語を使い，どのようなことができるのか直接的な情報を提供しなければならない。このためには，確固たる評価基準を設定する必要がある。現在は，主に評価指標と評価基準を定めて判定に使うルーブリック（Rubrics）の手法が評価に用いられる（Stevens and Antonia, 2013）。

代表的なルーブリックに「分析的評価法（analytic evaluation）」と「全体的評価法（holistic evaluation）」がある（Cohen，1994; Bachman and Palmer，1996）。

分析的評価法は，例えば語彙的能力，文法的能力，ディスコース能力などの異なる構成要素で評価し，それぞれの尺度で得点を与えるものである。この方法は，被験者の弱点などを見つけて診断を行うのに適している。しかし現段階では，ライティング能力の構成要因をいかに分けるのか，またそれぞれにどのような配点を与えるのか明確な実証的根拠がない。さらに，文書の全体的印象で，それぞれの評価項目に何らかの影響を与えかねない。日本人学習者が受験する TOEFL，IELTS，TEAP，G-TEC，英検などは，タスクの達成度と，語彙的能力，文法的能力，ディスコース能力のルーブリックに基づき，最終的に統合的に評価するものが多い。

一方，全体的評価法は，ライティング能力を統合的言語能力と考え，言語熟達度を１つの尺度で測る。このため，文書表現に関する弱点を診断するのには適していない。しかしながら，実生活の文書によるコミュニケーションでは，書き手の意図がどれだけ相手に伝わるかが重要で，構成要因に注意を払うことは少ない。また，この方法の利点として，評価基準が細分化されていないので採点しやすく，他の評価者との信頼性も高めやすい。また，実用的な面で実際の指導現場で導入しやすい（Weigle，2002）。

2.4　CEFR-Jのライティングの評価

現段階で CEFR-J のテストタスクの評価については，いずれの技能も統合的評価を活用している。これは CEFR-J がアクション・オリエンティッドで，実際の場面で何ができるかに焦点を当てており，タスクの達成度をより重視しているためである。また，各レベルのディスクリプタの内容が異なり，タスクによって求められる技能がそれぞれ違う。特に，現場の教員への活用という将来性の観点から統合的評価の方が実用的であろう。

根岸科研のライティング班では，テストタスクの構築方法や，評価基準

の選定に討議を繰り返し，予備検証を経て合意を得た[8]。

CEFR-J のライティングテストの評価は，それぞれ以下のような３段階の評価となっている。

評価１　未達成：そのレベルに達していない。

評価２　（最低限）達成：そのレベルに達している。

評価３　（余裕を持って）達成：次のレベルに達している可能性がある。

この３段階の統合評価は，CEFR のレベルとリンクされているCambridge English Exam においても同様に，Below Level，Pass with Merit，Pass with Distinctionの３レベルで最終統合評価として受験者に報告される[9]。

またこの３段階評価はタスクレベルによっての詳細な評価基準がある。例えば，CEFR-J ライティングタスクB1.2.2の評価基準はディスクリプタの Can-Do に合わせて以下のようになっている。

・B1.2.2の Can-Do

物事の順序に従って，旅行記や自分史，身近なエピソードなどの物語文を，いくつかのパラグラフで書くことができる。また，近況を詳しく伝える個人的な手紙を書くことができる。

テストタスクの条件

①３年前から変わったこと，②最近あった印象的な出来事が書かれている。③複数のパラグラフで順序だって報告している

評価1	・３つの観点について書かれていない ・文章にまとまりがない。 ・文法や語彙の間違いがあり文の意味は通じない
評価2	・３つの観点についてほぼ書いている ・文章にだいたいのまとまりがある。 ・文法や語彙の間違いはあるが文の意味は通じる ・語彙や文法構造に多少のバラエティーがある
評価3	・３つの観点を適切に書いている ・文章にまとまりがある。 ・文法や語彙の間違いはほとんどない ・語彙にバラエティーがあり，時に複雑な文法構造の文を使っている

以上のように，タスクの達成度を中心に，ディスコース能力として文章のまとまりを評価し，文法や語彙の適切さと多様さで総合的に3段階評価を行うものである。今回の予備調査では，A2.2.2，B1.1.2のテストタスクでも同様に，それぞれの Can-Do に合わせて構築した評価基準を採用した。

2.5　自動レベル判定システム

これまで多くの自動レベル判定システムが構築され実用化が図られている（Crossley，Roscoe，and McNamara，2013; Crossley，Kyle，and McNamara，2016）。前述のように代表的な物として ETS による *e-rater* などがある。Cushing，（2010）によると，*e-rater* による TOEFL iBT Independent Writing の採点結果を，手動による採点結果と比較し，中程度の相関関係が見られた。同様に Enright and Quinlan（2010）では，*e-rater* と手動判定結果に，一定の内容的妥当性や規準関連妥当性があるとされている。TOEFL のように大量の受験者がいるテストでは，ライティングテスト採点の複雑さや煩雑さに対処するには自動レベル判定の役割は大きい（Llosa and Malon，2018）。

CEFRに関連した自動レベル判定として，前述の Write & Improve がある。これは Cambridge English の同サイトで与えられたタスクを選択し，ライティングを入力すれば，レベル判定の結果と，フィードバックが得られる。特にこのサイトは無料で活用できるので教育現場での汎用性はある（齋藤，2017）。

しかし，自動レベル判定システムの問題点も様々指摘されている（Weige，2013）。自動システムでは，与えられたタスク特有のコンテクストや内容に基づく学習成果の把握は容易でない（Wagner，Foster，and van Genabith，2007）。また，必ずしも個別の被験者に対する詳細なフィードバックが得られるとは限らない。特に，既存の自動レベル判定システムを構築する基のデータは，日本語以外の様々な母語を持つ受験生のライティングを活用している。このため，日本の英語学習者特有のエラーの反映や，

それに基づくレベル判定やフィィードバックは考慮されていない。

以上の観点から，自動システムも日本の英語学習者のライティングデータやエラーに基づくシステムを活用する方がより実用的だと考えられる（e.g.，Matsuno，2009）。

2.6　CEFR-Jライティング自動レベル判定システム

このシステムは，前述のJEFLLコーパスを基に東京工業大学の奥村研究室が開発したものである。このコーパスは，日本の中学高・高等学校における約１万人分の英語学習者が書いた自由英作文データにより構築された。またこのコーパスは，英語学習者によって書かれた英作文の原文と，それを基に英語教育者が文法的に訂正した情報が含まれている。

一般に学習者のエラーをライティングのレベル判定に使う手法の整合性は認めらており，*e-rater* でも活用されている（Wagner，Foster，and Genabith，2007）。その主流は，全英作文内の誤りの出現割合を素性として活用している。これに対してCEFR-J 自動修正システムは，誤り訂正モデルを活用して，学習者がどのような誤りをしているかという傾向を基にレベルを推定する（Hayashi，Sasano，Takamura and Okumura，2017）。

被験者のライティングデータを入力することで，CEFRでのレベル，文法使用項目，単語数，文数，A1からB2までの語彙使用割合が抽出される（林・笹野・高村・奥村，2016）[10]。

３．研究

3.1　研究仮説

前章までの議論で明らかになったように，これまで日本人英語学習者の具体的な誤り傾向などを基にしたCEFRの自動レベル判定と，手動判定の比較を行った研究はない。このことから本研究では以下の仮説を立て，こ

の課題に取り組む。

・仮説１：CEFR-Jライティング・テストタスクの手動による判定と自動判定は，タスクの達成度を同様に評価する
・仮説２：CEFR-Jライティング・テストタスクの手動による判定と自動判定の結果には相関関係がある。
・仮説３：自動判定との比較を基に，タスクの内容や採点方法などの改善に有効な示唆が得られる。

3.2　ライティング・テストタスク検証の経過

CEFR-J ライティング・テストタスクのプロジェクトに関しては，これまで主に以下のような3段階の手順を踏んでいる。

3.2.1　初期予備テストPreA1-A2.1レベル

根岸科研ライティング班，工藤・能登原・中谷３名と，信州大学の酒井先生の協力で，PreA1.1からからA.2.1の５レベル各２つのディスクリプタ対応する10個のライティング・テストタスクを作成した。これを私立大学の大学生24名と，国立大学の付属校の中学生20名に実施し，順天堂大学小泉先生の協力を得て，各レベルの信頼性と妥当性の検証を行った[11)]。この結果を基にタスクや評価基準の見直しも実施した。

3.2.2　PreA1-A2.1レベル分析テスト

前述の改良を基に，関東の私立中学で２年生，３年生のべ157名にPreA1.1からA.2.1の５レベルのテストタスクの実施をした。それぞれのタスクの採点基準を統合的評価の３段階で構築した。これを基に各タスク２名の採点者が別々に採点を行い，結果を互いに照合し，各被験者の最終評価を行った。

3.2.3　A2.2-B1.2.2レベルのテストタスク構築

続く検証として，次のレベルであるCEFR-JのA2.2.1，A2.2.2，B1.1.1，B1.1.2，B1.2.1，B1.2.2の６レベルにおけるテストタスクの作成と評価基準の設定を目指した。テスト実施結果から得られる示唆を基に，これらのタスクの信頼性や妥当性を検証するのが最終的な目標である。

関東の私立高校２校A，Bにおいてのべ約205名の被験者にテストタスクを受験してもらった。この際，上の検証3.2.2との関連性の観点から1部の被験者には，アンカータスクとして，A2.1レベルも受験してもらった。

これらの回答を，3.2.2と同様に，採点基準を基に各タスク２名の採点者が別々に採点を行った。これらの結果を互いに照合し，各被験者の最終評価を実施した。

3.3　リサーチデザイン

本研究は，上記3.2.3節におけるA高校の１クラス37名に実施した，３レベルのタスクA2.2.1，B1.1.2，B1.2.2の回答と採点結果を基に，著者が行った予備調査の検証である。

3.3.1　データ収集

2019年２月にA高校でライティング・テストタスクを行った。本研究のデータは１つのクラス参加者37名にA2.2.1→B1.1.2→B1.2.2の順で実施したものである。各15分で回答してもらい，そのつど回収を行った。

3.3.2　採点手順

ライティング班が採点基準とマニュアルを作成し，採点のサンプルを事前に作成した。採点者は英語教育分野の大学院研究生２名である。マニュアルを基に採点基準を把握し，採点サンプルを確認した。ライティング班のメンバーが同席し，採点者が採点方法を習得するまで同時に採点を行った。その後，それぞれが各被験者の回答を３段階で採点し記録した。２者

の採点間で相違がある場合は，タスク班メンバーと共に回答を確認し最終的な合意を得た点数を記録した。前述のように，この中でA高校の１クラス37名が回答した CEFR-J の３レベルテスト結果を対象に今回の検証を行った。

3.3.3　自動レベル判定システムの入力と結果の記録

上記37名×３レベルの111の手書き回答を英語教育分野の大学院研究生に書き起こしをしてもらった。この際，エラーや改行などは，そのまま記録された。述べ111の学習者のコーパス・データを本論の検証に使用した。

この111のデータを筆者が CEFR-J 自動レベル判定システムに入力し結果を産出しエクセルシートに保存した。出力結果として，使用されている単語数，文数，文法項目数とその項目を記録した。また，各コーパスにおけるA1，A2，B1，B2の各レベルの語彙使用割合と，最終的なレベル判定も記録した。

3.4　分析方法

手動の判定結果は，それぞれ評価１を１点，評価２を２点，評価３を３点とした。自動レベル判定の方は，A1を１点，A2を２点，B1を３点，B2を４点とした。

3.4.1　手動及び自動レベル判定による各レベル到達人数の比較

各レベルの到達人数を比較するため以下のような手順を取った。はじめに，A2.2.1タスクで手動評価２以上の被験者数と，自動判定でA2以上の評価を得た被験者の数を比較した。次にB1.1.2タスクで手動評価２以上の被験者数と，自動判定でB1以上の評価を得た被験者の数を比較した。さらにB1.2.2タスクで手動評価２以上の被験者数と，自動判定でB1以上の評価を得た被験者の数を比較した。

3.4.2　手動及び自動判定の相関関係

各レベルのテストタスクの手動による点数と，自動評定の点数の相関関係をピアソンの相関関数で検証した。まず，各レベルで検証し，続いて全てのテストを合わせた結果の相関係数を計算した。

3.4.3　手動及び自動判定の相違の大きい回答の質的検証

手動及び自動判定の相違の大きい回答を選び，コーパス・データを確認した。書かれている内容を，評価基準のタスク達成度，語彙・文法等の適切さやバラエティーさの項目を参照に確認した。同様にエラーなども確認して，なぜ相違があったのかについて考察を行った。

４．結果

この章では仮説の検証の観点から結果を表示していく。尚，付表１に37名の３レベルのタスクにおける手動判定と，自動レベル判定の結果全てを掲載している。

4.1　仮説１の検証

「CEFR-Jライティング・テストタスクの手動による判定と自動判定は，タスクの達成度を同様に評価する」

表１に３つのタスクごとに手動による判定で２点以上あり，該当レベルに達した人数を掲載している。また同様に自動システムで該当レベルに達している人数も掲載している。さらに，両判定でそのレベルに達している，もしくは達していないという結果が一致した人数を掲載している。

手作業でA2.2.2レベルに達していると判断されたのは５人で，B1.1.2では10人，B1.2.1レベルで12人となっている。通常のテストでは，同じ被験者グループでは，レベルが上がるとタスクの難易度が上がり合格者の数は

減る傾向である。今回の手動判定では，レベルが上がることに合格者も増えるといった結果になっている。これは，各タスクの要求される難度の順番に課題があるか，採点基準に問題があったのかもしれない。この点は今回は検証していない。今後, 同時にテストを受験した他の被験者の結果や，他のレベルの結果を加えて，大きいサンプルを用いてラッシュモデル等でタスクの難易度などを確認する必要があろう。

一方，自動レベル判定システムでは，A2.2.2のテストでA2以上と判定されたのは32人であるB1.1.2テストでB1以上の判定は16人，B1.2.1のテストでB1以上と判定されたのは 4 人となっている。こちらの判定は，テストのレベルが上がるにつれ，合格者が減少するという通常のテスト結果に沿っている。

表 1 の右の欄は，前述のように両方の判定で共に該当レベルに達していると評価された人数と，共に達していないと評価された，結果の一致した人数を掲載している。これによるとA2.2.2のテストでは10名で, 全体の27%の被験者の結果に一致を見たことになる。同様に，B1.1.2テストで20人の54%が一致した。またB1.2.1のテストでは25人の68%が一致している。

このことからテストのレベルが上がるにつれ，手動採点と自動システムでは一致が見られる傾向が分かった。特にB1.2.1では， 7 割近くの被験者に同様に判定が行われた。このことによりCEFR-J の Can-Do の内容が高度になり，より複雑な英文産出を求められるテストでは，両方の結果に整合性が得られやすいのかもしれない。

表 1　手動採点と自動システムによる判定の比較

タスク	手動合格人数 2 点以上	自動システム合格人数	合格・不合格の 一致人数
A2.2.2	5 人	A2以上　32人	10人　(27%)
B1.1.2	10人	B1以上　16人	20人　(54%)
B1.2.1	12人	B1以上　4 人	25人　(68%)

結論として，手動による判定と自動判定は，タスクのレベルにより，目

標レベル達成度の評価には差がある。以上のことから，仮設１は必ずしも支持されたとは言えない。ただし，レベルが高いタスクの場合は，両者の判定に，ある程度同様な結果を得られることが分かった。

4.2　仮説２の検証

「CEFR-J ライティング・テストタスクの手作業による判定と自動判定の結果には相関関係がある」

表２に手動採点と，自動判定システムによる結果のピアソン相関の検定結果を掲載している。A2.2.2のタスク37名の手動採点による判定結果と，と自動判定システムの結果を数値化した値との相関係数は r =0.132となり，ほとんど相関がないと考えられる。つまり，手動判定の結果で高い点数を得た被験者が，自動判定結果で高い点数を得る傾向があるとは言えない。同様にB1.1.2の両者の相関係数は，r =0.073と低く，ほとんど相関がない。また，B1.2.1における両者の相関係数は r =-0.087となり，符号は負であるが，やはり負の相関係数があるとも言えない。さらに，手動と自動判定における３つのタスクのすべて得点を比較した場合も，相関係数は r =0.03と低く，ほとんど相関関係はなかった。

以上の結果から，手動採点と自動判定システムによる結果の相関関係があるとは言えず，仮設２は支持されなかった。これは，仮説１の検証において，タスクのレベルが上がると手動採点では合格者が増えたが，自動では合格者が減っていたという観点からも確認できる。

表２　手動採点と自動判定システムによる結果の相関関係

タスク	手動判定		自動判定		
Item	Av.	SD	Av	SD	Correlation
A2.2.2	1.189	0.518	2.297	0.777	0.132
B1.1.2	1.270	0.450	2.405	0.798	0.073
B1.2.2	1.324	0.475	1.595	0.762	-0.087
３タスク総合	1.261	0.481	2.099	0.852	0.003

Av.：平均，SD: 標準偏差

4.3　仮説３の検証

「自動判定との比較を基に，タスクの内容や採点方法などの改善に有効な示唆が得られる」

ここでは上の仮説を検証するために，手動採点の結果と自動判定システムの結果において，特に差の大きかった被験者の回答を確認する。付表１に示した，全得点の中から，両方の得点に３の差があったものを取り扱う。

表３には，手動採点と自動判定システムの得点差が３の事例を示している。111のデータの中で５つの事例が見られた。A2.2.2で２件，B1.1.2で２件，B1.2.2で１件であった。いずれも手動では１であったが，自動判定の得点は４であった。以下に，自動システムの優位な事例，両方で確認したほうが良い事例，手動判定の方が優位といった，３つの特徴的な事例を確認する。

表３　手動採点と自動判定システムによる結果の差の大きい事例

タスク	被験者番号	手動評価	自動判定	自動得点	得点差
A2.2.2	10	1	B2	4	3
	17	1	B2	4	3
B1.1.2	16	1	B2	4	3
	20	1	B2	4	3
B1.2.2	17	1	B2	4	3

4.3.1　A2.2.2レベルの事例

A2.2.2レベルの Can-Do は以下のようになっている。

「聞いたり読んだりした内容（生活や文化の紹介などの説明や物語）であれば，基礎的な日常生活語彙や表現を用いて，感想や意見などを短く書くことができる」

テストタスクの条件

日本語の敬称「～さん」の使い方について，①どのように感じるか自分の感想，②使いたいかどうかの意見，③なぜかの理由，について自分の感想や意見が簡潔に書かれてある。

評価１	・50語以下である。 ・①②③の内容に欠けているものがある。 ・①②③の内容はそろっているが，感想や意見がわかりにくい。 ・文法や語彙の間違いが見られ，一部は内容の理解に支障がある。
評価２	・①②③の内容は書かれており，感想や意見がある程度はわかる。 ・文法や語彙の間違いは見られるが，内容の理解に支障がない範囲である。
評価３	・感想や意見が①②③の３つ以上適切に書かれている。 ・文章にまとまりがある。 ・文法や語彙の間違いはほとんどない。

（1）被験者10：自動判定が優位な事例

a. 回答　手動判定１

I think the act is right. In Japan, some people use “san” after Japanese names. It is often used for respect people. So I feel happy when it is used, Also if you have a more respect person, you should use “sama” after their names. Maybe the person will be glad and you will be able to be spoken by the person. Finally, their relationship will be better than before. For these reasons, I recommend to use “san” after not only Japanese names but also English names!

（下線はタスクの達成項目，網掛けはエラーまたは意味が通じにくいと思われる箇所に著者が加えた）

b. 自動判定B2

・使用文数８，単語数87，文法使用項目数17

・文法使用項目；６人称代名詞所有格（my/our/your/her/their），13 不定冠詞，14 定冠詞，15 決定詞（some/any），32 副詞（頻度），38 比較級（優勢比較）（-er）（better等を含む），69 時制・相（未来），73 受

動態（現在），88 to不定詞（to DO），91 受動態のto不定詞，101 動詞＋to不定詞，121 助動詞類（be able to），139 助動詞類（should），141 助動詞類（will），150 等位接続詞，156 副詞節（if），162 従属接続詞 thatの省略（hope/know/think）

・語彙使用 ; A1レベル47.1%，A2レベル15.3%，B1レベル11.8%，その他の機能語25.9%

c. 考察

タスク達成項目は，２つ以上書かれており，ある程度文章にまとまりもある。エラーはあるが，文の意味は通じると考えられる。手動の採点で，１としたのはエラーの個所によるものかもしれないが，このケースでは評価２が適切と思われる。

自動システムの結果では，多様な文法項目が使用されており，他の事例に比べてA2やB1レベルの語彙の使用割合も高かった。このことが，自動レベル判定でB2という結果になったのかもしれない。自動判定では，例えば their relationship の個所はエラーと判定しないかもしれないが，意味的には our relationship の方がより適切であろう。このようにコンテクストに合わせた詳細な判定を，自動システムでどのように取り扱うかも検討する余地があるかもしれない。

d. まとめ

この事例は手動採点を見直した方がよいであろう。自動のシステムで確認することにより，より適切な採点結果を得られるという事例と考えられる。

（2）被験者17：自動判定と手動両方で確認が必要な例

a. 回答　手動判定１

I think I will feel a bit strange to use “san” after English names. In

England, I think people like not to use “san”. I don’t like to be used “san” after my name. I think to be friendly with a lot of people, not to use “san” after each names.
（下線はタスクの達成項目，網掛けはエラーまたは意味が通じにくいと思われる箇所に著者が加えた）

b. 自動判定B2

- 使用文数 4，単語数51，文法使用項目数10
- 文法使用項目；6 人称代名詞所有格（my/our/your/her/their），13 不定冠詞　69 時制・相（未来）　88 to不定詞（to DO）　89 to不定詞の否定（not to DO）　91 受動態のto不定詞　101 動詞+to不定詞　102 動詞+not+to不定詞　141 助動詞類（will）　162 従属接続詞 that の省略（hope/know/think）
- 語彙使用A1レベル47.6%，A2レベル16.7%，B1レベル9.5%，その他の機能語26.2%

c. 考察

回答の文数は 4 で，単語数51と短く条件の50語を何とかクリアしている。ただタスク達成項目は，2 つ以上書かれており，ある程度文章にまとまりもある。4 文目のローカルエラーの names の箇所は，意味は通じるが，等位接続詞が抜けている，あるいは構文的に的確な構造とは言えない。これらの点を考慮して，意味は通じにくいと手動の採点者は判断し，1 の評定をしたと推測される。

自動判定ではB2となったのは，網掛けで示した部分は，それだけを見ると文法項目として間違ってはいない。むしろ，文法項目における，89 to不定詞の否定（not to DO）や，91 受動態のto不定詞，101 動詞+to不定詞，102 動詞+not+to不定詞など複雑な構文が使用されていると自動判定されたと考えられる。一つの理由として，自動採点の基データの JEFLL はA1

からB1判定のデータが多く，重文・複文の構造における Grammar の整合性を指摘された誤り例が少ないからではないだろうか。最後の文のような，長い文全体における構造のエラーは，現状のバージョンでは手動採点で確認したほうが良い。

d. まとめ

これは，手動の採点結果の方が望ましいが，自動採点で確認することで，より適切なフィードバックができる事例と思われる。自動判定では，微妙な語用法や，重文・複文の構造の判定は容易でないが，４番目の文を少し改善すれば意味の通じる文を完成できる。CEFR の基準では，Aレベルにおいてタスクの達成度が重視され，この被験者はその点は到達している。このような事例の被験者には，該当箇所のエラーの改善を促すことで判定の向上が望めるだろう。

4.3.2　B1.1.2レベルの事例

B1.1.2の CAN-DO は以下のようになっている。

「身近な状況で使われる語彙•文法を用いれば，筋道を立てて，作業の手順などを示す説明文を書くことができる」

テストタスクの条件

台風接近による交通機関の乱れへの適切な対応手順について，①学校の方針（災害対策）を確認する，②台風情報（警報など）を確認する，③駅

評価１	• 災害時の対応手順がわからない。 • 文章にまとまりがない。 • 文法や語彙にバラエティーがほとんど見られない。
評価２	• 災害時の対応手順がある程度はわかる。 • 文章にだいたいまとまりがある。 • 語彙や文法構造に多少のバラエティーがある。
評価３	• 災害時の対応手順が適切に書かれている。 • 文章にまとまりがある。 • 文法や語彙の間違いはほとんどない。 • 語彙にバラエティーがあり，時に複雑な文法構造の文を使っている。

のアナウンス（運行状況）を確認する，など，場面で考えられる手順が自然な流れで書かれてある。

（3）被験者16：手動判定が優位な事例

a. 回答　手動判定 1

Do you get on the train now? If you get on the train, you wait there to move to train again. If you don't get on the train, you should go to the safety place. For example, high biluding, and school etc. If you can't go to safety place because it rains hard, you wait for the station, and when you can move, you should go to the safety place. And I think that you can above the bad situation, so you move relax.

（下線はタスクの達成項目，網掛けはエラーまたは意味が通じにくいと思われる箇所に著者が加えた）

b. 自動判定B2

- 使用文数 6，単語数84，文法使用項目数 8
- 文法使用項目；14 定冠詞　88 to不定詞（to DO）　101 動詞+to不定詞　123 助動詞類（can）　139 助動詞類（should）　150 等位接続詞　152 that 節（目的語）　161 従属節（as/if/that/when/whether以外の主な従属接続詞）
- 語彙使用A1レベル50%，A2レベル15.2%，B1レベル，9.1%その他の機能語25.8%

c. 考察

回答では，いくつかの文で手順を記述しているように見えるが，意味の通じない箇所が多く，手動の判定は 1 となったと思われる。to move to train againはuntil the train moves の意味で被験者は書いたのかもしれない。For example は前のピリオドをコンマにして，F を f にすれば改善されるとい

ったパンクチュエーションのエラーともみなせる。しかし，Grammar Profile では，パンクチュエーションの領域はカバーしていない。また，wait for 自体は動詞句として使われるが，この場合は wait at の方がコロケーション的に良いであろう。above は avoid と書こうとしたのかもしれないが，手動では動詞の欠如とみなされ意味が通じない文となる。move relax のも語用法として誤りとみなせる。被験者16は，特有のエラーで意味の通じない箇所が多い極端な事例で，手動で判定するしかないと思われる。

d. まとめ

この事例は，手動の判定の方が適切で，自動判定システムでは確認できない意味的なエラーが多い特徴的なものである。意味的に不適切なものや，ご用法の観点から自動判定システムの改善の可能性を示している。

以上の３つの事例より，仮説３の自動判定との比較を基に，タスクの内容や採点方法などの改善に有効な示唆が得られる点は支持されたと言える。

５．結論

今や英文ライティングの自動判定システムの開発が進み，実際のテストなどで活用が進んでいる。日本においてもセンター試験に代わり英語4技能試験の導入が予定されており，ライティング採点は手動では対処できず自動システムを導入するであろう。しかしながら，Weigle（2013）が指摘しているように，同じ語彙や文体でも，コンテスクによって意味や語用法の異なるものがあり，自動判定では困難なものがある。ところが，既存の開発者側からの研究発表の内容は，結果の信頼性や妥当性の報告は多いが，手動の判定の不一致があった場合の詳細な検討は少ない。

本論では，CEFR-J に即したライティング・テストタスクの開発の一環として，自動システムと手動採点の関連性を確認した。特に，両者の差が

大きい場合の事例のいくつかを詳しく確認した。

仮説１では，まず37名の３レベル合計111のライティングサンプルの判定を，手動と自動の各レベルの到達人数の比較を比較した。手動ではA22，B1.1.2，B1.2.1とレベルが上がるにつれ到達人数が増え，自動では減っていった。これは，同じ被験者が受験した場合，通常は自動判定のような結果を得ることが考えられる。このため手動の採点基準やタスクの難易度を見直す必要があるかもしれないことを示唆している。

続いて，両者の合格・不合格の一致した人数は，A2.2.2で27%，B1.1.2で54%，B1.2.1では25人の68%と，レベルが上がるにつれ，精度がよくなることが示唆された。これはタスクの課題達成要求が高まり，ライティングの複雑さや産出する量が増すことで，両者でより適合な判定が望めることを示唆している。

一方，仮説２の検証では，手作業による判定と自動判定の結果には，各タスクにおいても，３つのタスクの総合においても相関関係があるとは見なせなかった。この結果は仮説１の検証において，両者の合格判定人数がレベルで逆転していることからも推測できる。

仮説３では，特に両者で差が大きい場合の３つの事例を実際のライティングのデータを基に質的に確認した。自動システムが優っている例，受験者への適切なフィードバックのためには両方確認が必要な例，手動が優っている例である。このような検証は先行研究では十分行われていないので，不一致の起こる原因を考察できた意義は大きい。

これらの結果から，手動だけで学習集者のレベル判定を行うより，自動判定システムの結果と比較しながら，最終的な判定をする方が効果的であることが示された。また，学習者に納得のいくフィードバックを与える際にも，両者の比較から判定を行うことは有効である。

通常の指導現場では，学習者の書いたものを複数で採点するのは，時間的にも，労力的にも容易ではない。この際，自動判定システムの結果を参照にしながら成績やフィードバックを行うことは指導効果があると考えら

れる。

本稿は，CEFR-J ライティング・テストタスク構築のための予備調査である。今後，205人が受験した３タスクの合計615の書き起こしデータを活用した自動判定と，手動判定の結果を比較検証してくための調査方法構築の確認を目的としている。全体の約18%の111のデータを使い，判定結果の傾向や，両判定の整合性や相関関係などを調査し，残りの調査方法への示唆を得るための検証である。このため CEFR-J の３レベルの調査でサンプルも多くないため，この研究成果は断定的というものではない。また手動判定の採点者に各配点の理由を直接確認したわけではなく，採点基準に基づき類推している。今後は，各得点を与えた根拠を記録してもらうことも効果的であろう。

今回使用した自動判定システムの汎用性は高く，実際の教育現場での活用に有効である。今後，重文や複文など複雑な構造文や，パンクチュエーションの対処方法が改善すれば，より精度が上がると思われる。また，使用文法項目の CEFR レベル判定や，各レベル使用割合も同時に産出できれば教員にとってフィードバックが与えやすくなる。

全ての自動システムに該当する課題として，タスクの達成項目の判定や，コンテクストにおける意味の整合性の判定などは，現状ではかなり困難だと推測される。ライティングタスクは，それぞれ内容やコンテクストで達成度の要求が異なるため，場合によっては，タスクごとに判断基準の設定変更が必要となるであろう。

今後，様々な研究上での改善は必要であるが，CEFR-J に準拠したライティングテスト実施の際には，手動と自動判定の結果を照合した上で，学習者に評価結果やフィードバックを行っていくことが望ましい。

謝辞

本研究は根岸科研・科学研究費基盤研究（A）課題番号16H01935の研究助成に基づく成果報告である。玉川大学・工藤洋路先生，同志社大学・能

登原祥之先生と筆者3名によるライティング・テストタスクA2.2.1からB1.2.2レベルの開発の途中経過の中で，筆者が予備調査として行った検証の報告である。両２名の先生方にはタスク作成から本予備調査に至る経緯でたいへんお世話になっている。尚，これまでのライティング・テストタスクの作成や信頼性の検証には順天堂大学・小泉利恵先生，信州大学の酒井英樹先生のご協力を得た。またテストタスクの実施に当たっては，根岸先生や他の根岸科研の先生方，及び東京外国語大学の院生のご協力で実現した。東京工業大学の奥村先生からは，自動判定システムの活用法など様々なアドバイスをいただいた。ここに関係者の先生方に深く感謝の意を示します。尚，本論の内容に関する問題点などがある場合は全て著者の責任となる。

注

1）科学研究費基盤研究（A）『英語到達度指標 CEFR-J 準拠の Can-Do 指導タスクおよびテスト開発と公開』課題番号16H01935，2016-2019年度，研究代表者：根岸雅史

2）独立行政法人大学入試センターHP平成31年度大学入試センター試験実施結果の概要
https://www.dnc.ac.jp/center/shiken_jouhou/h31.html

3）Cambridge English の以下のURLに詳しい。
https://writeandimprove.com/

4）科学研究費基盤研究（A）『第二言語習得研究を基盤とする小，中，高，大の連携をはかる英語教育の先導的基礎研究』課題番号6202010，2008-2011年度，研究代表者：小池生夫

5）科学研究費基盤研究（A）『小，中，高，大の一貫する英語コミュニケーション能力の到達基準の策定とその検証』課題番号20242011，2008-2011年度，代表者：投野由紀夫

6）科学研究費基盤研究（A）『学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究』課題番号：24242017，2012-2015年度，研究代表者: 投野由紀夫

7）JEFLL コーパスは日本人EFL学習者（中学１年～高校３年）の英作文コーパスで約70万語。https://scnweb.japanknowledge.com/JEFLL2/

8）CEFR-J 2019 Symposium　2019 年 3 月 23 日配布資料。
9）*Cambridge English KEY for Handbook for teachers* を参照。
10）文法使用項目に関しては成城大学の石井先生が中心となって開発された CEFR-J Grammar Profileを基に検査が行われる。
11）中谷安男・工藤洋路・小泉利恵・能登原祥之・酒井英樹
「CEFR-J ライティングタスクの評価」CEFR-J 2018　Symposium 要綱
2018年 3 月17日～18日。於　成城大学開催

参考文献

Bachman, L. F., and Palmer, A. S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Cohen, A. D. (1994). *Assessing Language Ability in the Classroom.* Boston: Heinle & Heinle Publishers.

Crossley, S.A., Roscoe, R., and McNamara, D. (2013) Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In *FLAIRS 2013-Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference* (pp.208-213)

Crossley S.A., Kyle, K., and McNamara, D.S. (2016) The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48, 1227–1237.

Crusan, D. (2010) *Assessment in the Second Language Writing Classroom.* Michigan: University of Michigan Press.

Cushing, S. (2010) Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing* 27-3, 335-353.

Enright, M. K., and Quinlan, T. (2010) Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27-3, 317–334

林正頼，笹野遼平，高村大也，奥村学（2016）「誤りの傾向と文の容認性に着目した英作文のレベル判定」『情報処理学会第227回自然言語処理研究会』, 1-7.

Hayashi, M., Sasano, R., Takamura, H., and Okumura, M.(2017) Judging CEFR levels of English learner's essays based on error-typeidentification and text quality measures. *Proceeding of the 18th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing) 2017.*

小池生夫　他（2008）『第二言語習得研究を基盤とする小，中，高，大の連携をはかる英語教育の先導的基盤研究（平成16年度～18年度科学研究費補助金基盤研究A研究成果報告書）』

Llosa, L, and Malone, E. M. (2018) Comparability of students' writing performance on TOEFL iBT and in required university writing courses *Language Testing,* 36-2, 235-263.

Lu, X. (2017) Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, 34-4, 493-511.

Matsuno, S. (2009) Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, 26-1, 75-100.

水本篤（2008）「自由英作文における語彙の統計指標と評定者の総合的評価の関係」『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基づく客観的作文評価指標の検討』15-28.

Nakatani, Y. (2013) Investigating criterial features of EFL textbooks based on the CEFR. *Journal of International Scientific Publication: Educational Alternatives,* 11-2, 183-189.

Nakatani, Y. (2015) Effective oral presentations by business people in TED: Implications for developing CEFR can-do lists. *International Journal of Conceptions on Management and Social Sciences,* 3-4, 81-83.

Nakatani, Y. (2016) Exploring business communication strategies based on CEFR. *International Journal of Language, Literature and Linguistics,* 2-3, 86-89.

Nakatani, Y. (2017a) Exploring writing strategies for guiding readers: The use of metadiscourse in CEFR-based textbooks. *International Journal of Management and Applied Science Institute of Research and Journals,* 3-11, 14-17.

Nakatani, Y. (2017b) The applicability of emotional intelligence through CEFR towards enhancing cooperative teaching and self-learning in Japan. *WWA Journal,* 6, 18-30.

中谷安男（2010）「国際ビジネス英語到達目標に関するインタビュー調査－CEFR-J の質的検証への考察」『東京理科大学紀要（教養篇）』，42号，91-109.

中谷安男（2016）「CEFRの上位者のビジネスコミュニケーション・ストラテジーの検証：英語活用社員の調査」『国際ビジネスコミュニケーション学会年

報』75号, 13-31.
根岸雅史（2008）「CEFR の日本人学習者への適用可能性」『明海大学大学院応用言語学研究』No10, 45-54.
根岸雅史（2012）「CEFR 基準特性に基づくチェックリスト方式による英作文の採点可能性」『ARCLE REVIEW』 6 巻, 80-89.
齋藤雪絵（2017）「自動採点システムを使った英語ライティング学習」『立教大学ランゲージセンター紀要』38号, 63-74.
Stevens, D. & Levi, Antonia J. (2013). *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning. Sterling,* VA: Stylus Publishing.
van Ek and Trim (1990) *Threshold.* Cambridge: Cambridge University Press.
Wagner, J., Foster, J., and van Genabith, J. (2007) A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* 112–121.
Weigle, C.S. (2002) *Assessing Writing.* Cambridge: Cambridge University Press
Weigle S.C. (2010). Validation of automated scoring of TOEFL iBT tasks against non-test indicators of writing. *Language Testing,* 27-3, 335–353.
Weigle, S.C.（2013）English as a second language writing and automated essay evaluation. In Shermis, M., & Burstein, J. (Eds.), *Handbook of automated essay evaluation* (pp. 36–54). New York: Routledge.

付表１　手動による判定と自動レベル判定結果

タスク	A2.2.2			B1.1.2			B1.2.2		
被験者番号	最終評価	自動判定	自動得点	最終評価	自動判定	自動得点	最終評価	自動判定	自動得点
1	1	B1	3	2	A2	2	2	A2	2
2	1	A1	1	2	A2	2	2	A1	1
3	1	A2	2	1	A2	2	2	B1	3
4	1	A1	1	2	A2	2	1	A2	2
5	1	A2	2	1	A2	2	1	A1	1
6	1	A2	2	2	B1	3	2	A1	1
7	2	A2	2	1	A2	2	1	A2	2
8	1	A2	2	1	A1	1	1	A1	1
9	1	A2	2	1	A2	2	2	A1	1
10	1	B2	4	1	B1	3	1	A1	1
11	1	A2	2	1	B1	3	1	A2	2
12	1	B1	3	1	B1	3	1	A2	2
13	1	B1	3	1	B1	3	2	A2	2
14	1	A2	2	1	A2	2	1	A1	1
15	2	A2	2	1	A2	2	1	A2	2
16	1	A2	2	1	B2	4	1	B1	3
17	1	B2	4	1	A1	1	1	B2	4
18	1	A2	2	2	B1	3	2	A1	1
19	1	B1	3	1	B1	3	2	A2	2
20	1	B1	3	1	B2	4	1	A1	1
21	1	B1	3	1	B1	3	1	A1	1
22	1	A2	2	2	B1	3	2	A2	2
23	1	A2	2	1	A2	2	2	A1	1
24	1	A2	2	1	B2	3	1	B1	3
25	3	B1	3	2	A2	2	2	A1	1
26	1	A1	1	1	A1	1	1	A2	2
27	1	B1	3	2	B1	3	1	A1	1
28	3	B1	3	1	A2	2	1	A2	2
29	1	A2	2	1	A2	2	1	A1	1
30	1	B1	3	2	B2	4	1	A2	2
31	1	B1	3	1	A2	2	1	A1	1
32	1	B1	3	1	A2	2	1	A1	1
33	1	A2	2	1	A2	2	1	A1	1
34	1	A1	1	1	B1	3	2	A1	1
35	1	A1	1	1	B1	3	1	A1	1
36	2	A2	2	2	A1	1	1	A1	1
37	1	A2	2	1	A2	2	1	A2	2
Total	44		85	47		89	49		59
Av	1.189189		2.2973	1.27027		2.40541	1.32432		1.59459
SD	0.518429		0.77692	0.45023		0.7979	0.47458		0.76229

※最終評価：採点者２名の最終評価，自動得点：自動レベル判定を得点に換算したもの，Total：合計，Av：平均，SD：標準偏差

A Comparative Evaluation of Human Raters' Approaches to the Automatic Level Judging System: A Pilot Study for Developing CEFR-J Writing Test Tasks and Assessment Methods

Yasuo NAKATANI

《Abstract》

This paper explores the relationship between the results of the automated scoring system based on CEFR-J and human raters' assessments. As a pilot study for further investigation dealing with more subjects, this study examines 3 different levels of CEFR-J writing test tasks for 37 participants. First, two independent raters evaluated a total of 111 test samples by using the CEFR-J assessment guidelines for each individual level. These results were compared with the assessment of a CEFR-J automated level judging system that utilized leveraged error types and text quality measures. The results show that although the indicators used for correlation are low, the consistency between each method of evaluation tends to be better at a higher level: B1.2.1. The qualitative analysis of the test samples with large discrepancies indicates that it is effective to use both human raters and methods and the automated level judging system when deciding candidates' final scores and giving feedback on results.