様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ５ 月 １４ 日現在

機関番号：３２６８３
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ０２９８１

研究課題名（和文）タスクに基づくライティングテストにおける自動評価採点システムの実用化開発

研究課題名（英文）Development of Automated Essay-Scoring System for a Task-Based Writing Test

研究代表者
杉田　由仁（Sugita, Yoshihito）

明治学院大学・文学部・教授

研究者番号：７０３６３８８５

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：本研究では、日本人英語学習者のライティング能力の推定により有効な自動評価採点システムを開発するために、1) Accuracy タスクの「言語的正確さ」の評価において、語彙や文法、スペル句読法などにおける誤りを特定し、統計指標化する方法を考案すること、2) Communicability タスクの「情報伝達効果」の評価において、課題との関連性を判定し統計指標化する方法を考案すること、3) 重回帰分析によって総合的評価を予測する回帰式を作成してその有用性検討として、サンプル数を増やして信頼性・妥当性の検証を行うことを研究目的として取り組んだ。

研究成果の学術的意義や社会的意義
Accuracy、 Communicabilityの予測得点とCriterion スコアとの相関係数により、基準関連妥当性の検証を行ったところ比較的強い相関が見られ、また両者の合計点をTBWT によって測定されるライティング知識体系の総和(総合評価）と位置づけて相関分析を行ったところ、強い相関が見られた。これらの結果から、自動採点結果はCriterion のパフォーマンスを一定程度反映していることが確認された。さらに、評価結果に対するアンケート調査により結果妥当性の検証を試みたところ、評価結果が利害関係者である高校生に与える影響は適切であったことが確認された。

研究成果の概要（英文）：This study focuses on the reliability and validity of an automated essay-scoring system for a task-based writing test. The system was revised and 150 second-year high school students participated in the trial of the system and the Accuracy and Communicability values were calculated by the resulting formulas. To estimate the degree to which the indices were collectively related to the prediction of the scores, correlation analyses were conducted. The results showed that moderately high correlation existed between the scores of the both tasks and their indices. To validate the predictions of the formulas, the values were compared to the Criterion scores of the high school students. Correlations between the Accuracy and Communicability values were significant. The composite values of the two tasks were also significantly correlated. The results of the questionnaire evaluating scores of the tasks showed that the scores were acceptable and appropriate for the students.

研究分野：英語教育学

キーワード：ライティング　自動評価採点

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

今世紀に入り、教育測定の分野において最も精力的に研究が行われてきた研究の一つが、小論文やエッセイの自動評価および採点の研究であると言われる (石岡、 2008)。最近 15 年間に開発・実用化されてきた英文エッセイ自動評価採点システムの代表的なものとしては、Educational Testing Service (ETS) の Electric Essay Rater (e-rater)、Vantage Learning 社の IntelliMetric、Knowledge Analysis Technologies (KAT) 社の Intelligent Essay Assessor (IEA) などがある。日本においても、大学入試等で出題される日本語小論文をコンピューターで自動採点するシステム (Japanese Essay Scoring System: JESS) の開発が大学入試センターの研究者を中心に進められており、近い将来における入試等への本格導入なども想定される。

本研究では、「タスクに基づくライティングテスト(Task-based Writing Test、 以後 TBWT と呼ぶ)」に特化した「自動評価採点システムの実用化開発」を最終目的として着手した。平成 19～21 年度の第一次研究において、第二言語知識体系の二層性 (規則に基づく体系と記憶に基づく体系) に着目し、この二層性の体系間におけるトレードオフの関係をテスト法の原理とする言語能力測定モデル(Construct-based Processing Approach to Testing) を考案した (Sugita, 2009a)。

平成 22～24 年度の第二次研究においては、科学研究費による補助を受け「タスクに基づくライティングテストの開発とその有用性に関する総合的評価」という研究テーマに取り組んだ。第一次研究の成果である言語能力測定モデルをベースとして TBWT を開発し、実施方法、評価基準、評点の意味づけなど実施形態についての検討および実際に行ったテスト結果からテストとしての信頼性、妥当性、実用性についての検証作業を行った。その結果として、(1) “TBWT Scoring Guide” および評定尺度の最適化を図ることができた (Sugita, 2009b)、(2) ライティングテストとしての信頼性・妥当性の向上を図ることができた (Sugita, 2010)、(3) 評価基準は評定者を妥当な評価行動に導くものであり実用性が確認された (杉田, 2012)、(4) 未熟な評定者であっても、事前トレーニングにより、評定の一貫性向上が図られることが確かめられた (Sugita, 2012)、(5) 応用的研究の準備段階として、同等の困難さを持つ評価タスクを考案することができた (Sugita, 2013)。

さらに、平成 25～27 年度の第三次研究においても、科学研究費による補助を受け「タスクに基づくライティングテストにおける自動評価採点システムの開発」というテーマで、一定規模の本格的な公開テストの実施に向けて、信頼性および実用性の高いライティングテストの自動評価採点システムの構築に着手した。具体的には、TBWT を有用性の高いライティング・パフォーマンス・テストとして規模拡大を図るために、受験者の増加に対応し得る迅速かつ客観的な評価採点方法を整備することを目的として、近年著しい進歩を遂げている「コンピューターによる自動評価採点システム」を TBWT に特化して応用開発し、より客観的で実用的な評価システム (TBWT 自動評価採点システム) の構築に向けて取り組んだ。

応用的研究の内容として、(1) 英文における自動評価採点システムに関する研究動向の把握、(2) TBWT における言語的特徴数値化のための統計指標の選定とその検証、(3) コンピーター利用による自動採点システム処理系統について検討を行った結果、下記①～④ が実用化に向けての最終的な課題として浮かび上がってきた。

① TBWT の評価基準 (評価の観点) に適合する言語的特徴として抽出された客観的評価指標 (特徴量) により、総合評価を 61～69%予測できる回帰式を作成することができたが、予測精度をより向上させる可能性のある統計指標について再検討すること。

② 自動評価採点システムによる最終的な総合評価には、各評価指標に基づく「数値データによる予測式」方が３段階評価によるそれよりも有用性が高いことが示唆されたが、サンプル数を増やして実証すること。

③ Accuracy タスクの「言語的正確さ」の評価において、語彙や文法、スペル句読法などにおける誤りを特定し、統計指標化する方法の再検討および再設定を試みること。

④ Communicability タスクの「情報伝達効果」の評価において、課題との関連性を判定し統計指標化する方法の再検討および再設定を試みること。

研究の最終目的とする「TBWT 自動評価採点システムの実用化開発」に際しては、まず上記の課題解決に取り組み、システムそのものを完成させた後、パフォーマンス評価の信頼性・実用性を検証するために一定規模の調査を実施して実用化開発の達成を図ることにした。

２．研究の目的

本研究の目的は、TBWT に特化した「コンピューターによる自動評価採点システム」の実用化に向けて、最終的に残された課題を解決してシステムを完成させ、コンピューター利用によるパフォーマンス評価の信頼性・実用性を検証するために一定規模の調査実施を目指すことである。具体的には、(1) Accuracy タスクの「言語的正確さ」の評価において、語彙や文法、スペル句読法などにおける誤りを特定し、統計指標化する方法を考案すること、(2) Communicability タスクの「情報伝達効果」の評価において、課題との関連性を判定し統計指標化する方法を考案すること、(3) 重回帰分析によって総合的評価を予測する回帰式を作成してその有用性検討として、サンプル数を増やして信頼性・妥当性の検証を行うことを目的とする。

３．研究の方法

(1) 自動採点システムの改修

研究目的(1) に関しては、Accuracy タスクの「言語的正確さ」の特徴量として、入力された英文サンプルの総単語数に対するスペルに誤りがある単語数の割合 (以後、単語誤り率と呼ぶ) を追加設定する。フリーソフトウェアのスペルチェッカー Ispell の辞書 (https://lasr.cs.ucla.edu/fmg-members/geoff/ispell-dictionaries.html) に掲載されている 354、984 語に相澤・石川・村田 (2013) による JACET 8000 と Plus 250 およびその活用形をテキストファイルで追加・合成し、照合用のデータベースを構築する。そのデータベースとサンプルとの照合により検出されるスペルミスの数を入力された総単語数で割って誤りがある単語の出現率を算出するロジックを実装する。

Communicability タスクの「情報伝達量」の評価指標については、これまでの定型表現頻度語数を特徴量として継続使用する。ただし、研究目的(2) に関わり、フリーソフトウェアのコンコーダンサーである AntConc (Anthony, 2016) を活用して分析し、課題との関連性において適切であると判断される定型表現 (to+動詞) を追加して照合用リストを改訂する。具体的には、2011 年に 31 名、2015 年に 14 名の日本人大学生によって入力され TBWT に蓄積された Communicability タスクのサンプルから 463 チャンクを取り出して、これまでのリストに掲載されていた 46 チャンクに未掲載かつ有意味な 50 チャンクを追加してリスト掲載数を倍増させる。

(2) 分析対象

システム改修のための分析対象としたライティング・サンプルは、Sugita (2016) と同様に 2008 年前期に 20 名の日本人大学生を対象として行った調査において収集され、Sugita (2010) および Sugita (2012) において延べ 15 名の日本人中学校英語教師により評定が行われた 40 サンプル (Accuracy タスク 20、 Communicability タスク 20) である。2010 調査および 2012 調査ともに、評定データの分析は FACETS, Version 3.63 プログラム (Linacre, 2008) を使用して行われた。また、システム改修後の自動採点による評価の信頼性・妥当性を検証するためのデータとして，2018 年 2～3 月に 150 名の高校 2 年生が受験した TBWT 300 サンプル (Accuracy, Communicability タスク各 150) の総合評価および同時期に受験を依頼した *Criterion* スコアを用いることにした。*Criterion* の課題については、Accuracy, Communicability タスクのトピック・内容構成との共通性を考慮して、“Day with a friend: Imagine that a good friend is visiting you for a day. Write an essay explaining how the two of you might best spend the day together.” というプロンプトを選定した。

(3) 分析手法

TBWT の構成概念である Accuracy, Communicability それぞれの評価規準に適合する言語的特徴数値化のために、改修後のシステムに設定された客観的評価指標は下表に示す通りである。Accuracy タスクについては、各サンプルの評定値、文章構成力の客観的評価指標として抽出した「語数」「平均文長」および言語的正確さの指標である「難語割合」「接続語句数」に「単語誤り率」を加えて変数間の相関分析を行う。また、Communicability タスクについては、各サンプルの評定値、伝達内容の質の指標である「平均文長」「難語割合」、情報伝達効果の指標である「アイディアの数」「頻度語数」を用いて同様の分析を行う。この内,「頻度語数」に関しては、システムの照合用リストを改訂版に差し替えて自動採点を行った結果を適用する。

TBWT の客観的評価指標

構成概念	Accuracy		Communicability	
評価規準	文章構成力	言語的正確さ	伝達内容の質	情報伝達の効果
評価指標 (特徴量)	総語数 平均文長	難語割合 接続語句数 単語誤り率	平均文長 難語割合	アイディアの数 定型表現の頻度語数

４．研究成果

本研究では、日本人英語学習者のライティング能力の推定により有効な自動評価採点システムを開発するために、(1) Accuracy タスクの「言語的正確さ」の評価において、語彙や文法、スペル句読法などにおける誤りを特定し、統計指標化する方法を考案すること、(2) Communicability タスクの「情報伝達効果」の評価において、課題との関連性を判定し統計指標化する方法を考案すること、(3) 重回帰分析によって総合的評価を予測する回帰式を作成してその有用性検討として、サンプル数を増やして信頼性・妥当性の検証を行うことを研究目的として取り組んだ。その成果と今後の課題は、下記①～④の通りである。

① Accuracy タスクの「言語的正確さ」の特徴量として、入力された英文サンプルの総単語数に対するスペルに誤りがある単語数の割合 (単語誤り率) を追加設定した。その結果、単語誤り率が低いサンプルは Accuracy 評価が高く、誤り率が高いサンプルは評価が低いという関係性があることが確認され、これを含む４指標による Accuracy 評価の説明率は、これまでのシステムにおける 69% を上回る 72%となり、測定精度の向上が認められた。

② Communicability タスクの「情報伝達量」の評価指標については、課題との関連性を定型

表現 (to+動詞) によって照合するためのリストを改訂し、これまでの定型表現頻度語数を特徴量として継続使用した。Communicability 評価の説明率は改修前と同じ 61%であることが確認されたので、システムとしての安定性を新たな予測式によって確かめることにした。

③ 150 名の高校２年生が受験した TBWT 300 サンプル (Accuracy、 Communicability タスク各 150) により、予測得点の信頼性・妥当性の検証を試みた。Accuracy 評価の信頼性に関しては、予測式によって算出された予測得点と語数、難語割合には強い相関が、平均文長とは比較的強い相関が認められた。これら３指標との相関係数は母集団でも意味のある相関係数として判断できることがわかり、予測得点の信頼性が確認された。また、Communicability 評価の信頼性については、アイディア数と頻度語数には強い相関が、難語割合とは比較的強い相関が、平均文長とは弱い相関が認められた。予測得点との無相関検定の結果、Communicability に関しては４指標と予測得点の信頼性およびシステムの安定性が確認された。なお、Accuracy 評価の新指標として導入した単語誤り率に関しては、二次分析により、追検証を行う必要性があることが示唆された。統計的指標として、測定精度の再検証および向上を図るための継続研究に取り組む必要がある。

④ Accuracy、Communicability の回帰式による予測得点と *Criterion* スコアとの相関係数により、基準関連妥当性の検証を行った結果、Accuracy と Communicability 評価の予測得点とは比較的強い相関が見られた。また、両者の合計点を TBWT によって測定されるライティング知識体系の総和 (総合評価) と位置づけて相関分析を行ったところ、強い相関が見られた。これらの結果から、回帰式によるタスク評価得点は *Criterion* のパフォーマンスを一定程度反映しており、基準関連妥当性が確認された。さらに、評価結果に対するアンケート調査により結果妥当性の検証を試みたところ、Accuracy 評価に関しては、89.1%が「とても適切である」または「適切である」と回答しており、評価結果が利害関係者である高校生に与える影響は適切であったことが確認された。Communicability 評価に関しても、77.5%が「とても適切である」または「適切である」と回答しており、高校生に与える影響は概ね適切であったことが確認された。しかし、今後の課題として、Communicability タスクの評価結果に対しては「頻度語数」の意味合いや、それにより内容に関わる評価が行われていることなどが高校生には伝わりにくいため、他の評価観点も含めて、結果の解釈がしやすくなるような観点名やそのフィードバック方法について再検討の必要性が認められた。

【引用文献】

①Anthony, L. (2016). AntConc (Version 3.5.6) [Corpus analysis toolkit]. Retrieved from http://www.laurenceanthony.net/software/antconc/

② Sugita, Y. (2009a). The development and implementation of task-based writing performance assessment for Japanese learners of English. Journal of Pan-Pacific Association of Applied Linguistics, 13(2), 77–103.

③Sugita, Y. (2009b). Developing and improving rating scales for a task-based writing performance test. JLTA Journal, 12, 85–103.

④ Sugita, Y. (2010). Reliability and validity of a task-based writing performance assessment for Japanese learners of English. JLTA Journal, 13, 21–40.

⑤Sugita, Y. (2012). Effects of rater training on raters' severity, consistence, and biased Interactions in a task-based writing assessment. JLTA Journal, 15, 61-80.

⑥Sugita, Y. (2013). Comparability of accuracy and communicability tasks: Are they all equally difficult? JLTA Journal, 16, 65-86.

⑦Sugita, Y. (2016). Examination of objective rating indices to an automated essay-scoring system for task-based writing tests. *ARELE, 27*, 17–32.

⑧相澤一美・石川慎一郎・村田年. (編著). (2013). JACET 8000 英単語. 東京: 桐原書店.

⑨石岡恒憲 (2008). 小論文およびエッセイの自動評価採点における研究動向, 『人工知能学会誌』23 巻 1 号, 17-24.

⑩杉田由仁 (2012). ライティング評価における評定者の行動分析と評価基準の妥当性検証, 『JACET 関東支部学会誌 (JACET-KANTO Journal)』No.8, 14-26.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Yoshihito SUGITA	４．巻 30
２．論文標題 Reliability and Validity of an Automated Essay-Scoring System for a Task-Based Writing Test	５．発行年 2019年
３．雑誌名 全国英語教育学会紀要（ARELE）	６．最初と最後の頁 17-32
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 杉田由仁　石井雄隆	４．巻 1
２．論文標題 タスクに基づくライティングテスト自動採点システムにおける客観的評価指標の検討(2)	５．発行年 2017年
３．雑誌名 第43回全国英語教育学会島根研究大会発表予稿集	６．最初と最後の頁 178-179
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yoshihito Sugita and Yutaka Ishii	４．巻 最終号
２．論文標題 Validation of an Automated Essay-Scoring System for Task-Based Writing Tests	５．発行年 2018年
３．雑誌名 Information Communication Technology Practice and Research 2016	６．最初と最後の頁 13-20
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 杉田由仁
２．発表標題 タスクに基づくライティング・テスト自動評価採点システム信頼性・妥当性の検証
３．学会等名 第44回全国英語教育学会京都研究大会
４．発表年 2018年

１．発表者名
杉田由仁　石井雄隆

２．発表標題
タスクに基づくライティングテスト自動採点システムにおける客観的評価指標の検討(2)

３．学会等名
第43回全国英語教育学会島根研究大会

４．発表年
2017年

１．発表者名
Yoshihito Sugita

２．発表標題
Objective rating indices to automated essay-scoring systems for writing assessment

３．学会等名
The 50th IATEFL Annual Conference（国際学会）

４．発表年
2016年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
明治学院大学英語教授法(TESOL)研究室
タスクに基づくライティング・テスト自動採点サイト
http://www.meijigakuin.ac.jp/~ysugita/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	石井　雄隆 (Ishii Yutaka) (90756545)	千葉大学・教育学部・助教 (12501)	2017年度末をもって研究分担者を退任