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

平成　２９　年　　６　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：　2013～2016
課題番号：　２５３７０７２３

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開発

研究課題名（英文）Development of an automated essay-scoring system for task-based writing tests

研究代表者
杉田　由仁（SUGITA, Yoshihito）

明治学院大学・文学部・准教授

研究者番号：７０３６３８８５

交付決定額（研究期間全体）：（直接経費）　　1,100,000円

研究成果の概要（和文）：本研究では、タスクに基づくライティングテストに特化した「コンピューターによる自動評価採点システム」の開発を進めた。ライティング評価を予測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。しかし、予測精度をより向上させるために、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必要性が示唆された。

研究成果の概要（英文）：This study focused on the validity of objective rating indices to an automated essay-scoring system for task-based writing tests. The relationship between the holistic ratings of Accuracy and Communicability tasks and their objective indices was examined. Independent variables to measure the holistic ratings were chosen from existing objective indices based on the constructs of each task: organizational skills, linguistic accuracy, communicative quality and effects. As a result of the correlation analyses, a moderately high correlation was found to exist between the holistic ratings of the Accuracy task and its indices (tokens and type levels). A moderately high correlation was also found between the holistic ratings of the Communicability task and its indices (type levels and relevant ideas). Through multiple regression analysis, the scores of each task could be predicted with moderately high dimensional accuracy.

研究分野：　英語教育学

キーワード：　教育評価・測定　ライティング　自動採点

１．研究開始当初の背景

第 2 次研究 (2010～2012) の成果として、「タスクに基づくライティング・テスト (Task-based writing tests: TBWT)」において(1) 評定者が一貫した評定を行うことのできる評定尺度および『評定の手引き』を開発することができた (Sugita, 2009a, 2009b)、 (2) 受験者のライティングにおける運用能力を測定する上で、信頼性・妥当性のある評価タスク開発を可能にする「構成概念に基づく言語処理的テスト法」の考え方を確立することができた (Sugita, 2010, 2013)、(3) 評価対象となるライティング能力を効果的に測定することのできる５段階の単特性に基づく評定尺度を開発することができた(Sugita, 2012a, 2012b)。

しかし、これまでに TBWT の採点に関わった評定者を対象として実施した「評定作業に関するアンケート」では、評定作業には疲労が伴い、大きな負担を感じることなく評定を行うことができるのは「20～40 名」程度という回答が多かった。また、第 2 次研究において、事前のトレーニングにより評定者の熟練度を高めることが可能であることは確かめられたが、実施規模拡大に向けて多数の熟練した評定者を育成するためには、相当の費用と時間を要すると考えられる。今後、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は、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しておかなければ、拡大化そのものが困難になることが予想された。

２．研究の目的

本研究は、応用的研究段階に位置づけられるものであり、一定規模の本格的な公開テストの実施に向けて、客観性および有用性の高い TBWT の評価採点システムを構築することを目的として行う。第 1～2 次研究において開発を行ってきた 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」を応用開発し、評定者要因による影響を除外したより客観的で実用的なライティング・パフォーマンス評価の実現を目ざす。

３．研究の方法

(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

最近 15 年間、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 (石岡, 2008) と言われるが、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」を構想するための基礎研究として、これまでに開発が行われてきた、特に英文における自動評価採点システムに関する文献研究に取り組んだ。具体的には、Educational Testing Service (ETS) によって開発された Electric Essay Rater (e-rater)、Vantage Learning 社によって開発された IntelliMetric、Knowledge Analysis Technologies (KAT) 社が所管する Intelligent Essay Assessor (IEA) などを研究対象として、それぞれの英文分析ツールとしての特徴を重点的に検討した。

(2) TBWT における言語的特徴数値化のための統計指標設定

これまでの自動採点システム開発において用いられてきた説明変数 (言語的特徴) として、単語の出現頻度ベクトル、文や句の長さ、代名詞や助動詞の数、議論を深めるための手がかり語や修辞句など (石岡, 2004)、流暢さを表す総語数、統語的複雑さを表す文あたりの従属節数と平均文長、談話的特徴を表す文あたりの接続語句数など (杉浦, 2008) の例を参照した。TBWT 自動評価採点システムにおける説明変数を抽出するために、過去に大学生英語学習者を対象として実施してきた TBWT における総合的評価決定の要因となった言語的特徴の分析を行った。具体的な分析方法として、これまでの自動採点システム開発において用いられてきた客観的評価指標 (特徴量) を吟味して、「書く」領域における言語運用能力の構成概念として想定した Accuracy, Communicability それぞれについて「統計指標」を設定した。次に、延べ 20 名の中学校・高等学校教員が評定を行ってきた TBWT の総合的評価と相関の強い指標を抽出し、重回帰分析により、総合的評価を予測する回帰式を作成し、その妥当性を検証した。

(3) コンピーター利用による自動採点システム構築に向けての検討

自動評価採点システムの構築に向けて、テキスト処理に使用するプログラムの開発に取り組んだ。プログラミング言語としては、Perl (Practical Extraction and Report Language) を使用した。習熟困難な部分もあったので、第一段階として Perl の基礎から理解を深め、プログラム作成の準備を行った。また、Web 上における受験を可能にするためのサーバーへのアップロード、テスト結果を管理するためのセキュリティ等に関しては専門の業者に依頼してソフト・ハード面における実施準備も並行して行った。

(4) 自動評価採点システムの信頼性に関する検証

入力されたテキスト処理のために開発された採点・評価プログラムの信頼性に関する検証を行った。検証方法としては、過去に実施された TBWT における評定者 (中学校・高等学校教員) の評価と自動評価採点システムによる評価との相関の度合いを信頼性の指標とした。

(5) 実用性に関する検証

TBWT 自動評価採点システムの教育現場における実用性について検証を行うために、

大学生学習者を対象として TBWT を実施した。対象となった学生には、妥当性検証の基礎データとなる「オンライン自動採点システム(Criterion)」によるエッセイライティングのテストの受験機会を設定し、客観的な評価指標に基づいて与えられる５段階の全体的評価法のスコアと TBWT 自動評価採点システムの評価結果との相関から併存妥当性について検討を行った。

４．研究成果と課題

今次の研究において抽出した客観的評価指標の妥当性を確認するために、それぞれのタスクの変数間の相関分析を行った結果は下記の通りである。

(1) Accuracy タスク

Accuracy 評価との相関を見ると「語数」との相関が r = .88 と高かった。特に Accuracy タスクは「120 語程度」という語数指定が行われていることもあり、この客観的指標と評定者による Accuracy タスクの観点別評価 (文章構成力) との相関が強いと考えられる。これに対して「平均文長」とは相関が見られず、語数の多い長めの文を書いても評定者による Accuracy 評価にはそれほど影響しないことが確認された。また、「言語的正確さ」の観点から客観的指標として設定した「難語割合」「接続語句数」はいずれも Accuracy 評価と中程度以上の相関が見られた。すなわち、JACET 8000 のレベルの高い語や文と文のつながりを示す接続語句の使用は、評定者による評価に影響を与えることが確認された。

(2) Communicability タスク

Communicability 評価との相関を見ると「頻度語数」との相関が r = .83 と高かった。Communicability タスクは定型表現 (to+動詞) を使用してアイディアを書くという言語形式の指定が行われているので、この客観的指標は評定者が「情報伝達の効果」の観点から Communicability 評価を行う際に影響が大きいと考えられる。また「アイディア数」に関しても中程度の相関が見られた。一方、「伝達内容の質」に関わる客観的指標については、平均文長 (r = .59), 難語割合 (r = .75) いずれも Communicability 評価と中程度以上の相関が見られた。特に、Accuracy タスクの場合には「平均文長」と相関が見られなかったが、一行に１つのアイディアを書く形式となっている Communicability タスクについては、語数の多い長めの回答を書くことが評定者による評価にも影響する可能性があることが確認された。さらに、JACET 8000 のレベルの高い語を使用することは「伝達内容の質」向上につながり、評定者による Communicability 評価に影響を与えることが確認された。

上記の妥当性検証結果に基づき、抽出された客観的評価指標による TBWT 自動評価採点の処理系統を図示すると、下図の通りとなる。

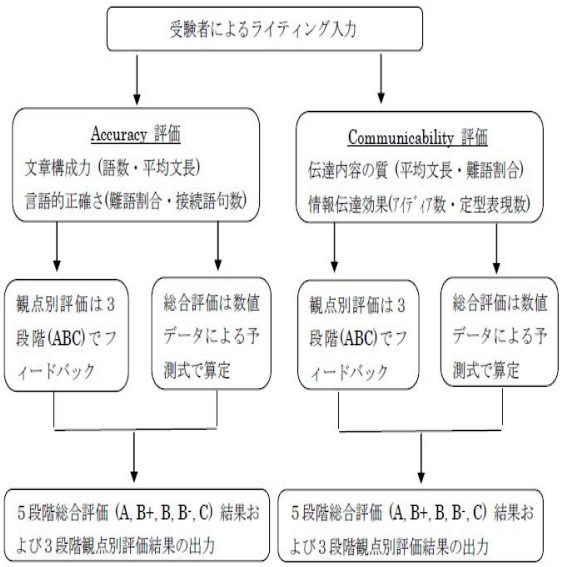

図　自動評価採点システムの処理系統

まず、受験者が入力を行った Accuracy タスクのライティング・サンプルに関しては、総語数および含まれる文章の数がカウントされた後、１文あたりの単語数 (平均文長) が算出される。また、JACET 8000 における難しい語の使用割合 (難語割合) については、レベル別カバー率分析プログラム (V8an. pl) によりレポートされる各レベルの見出し語の出現数にレベルを示す数値 (1～8) を掛け合わせる。接続語句数は、接続語リストに掲載されている語句の出現数による。それぞれの指標に対しては四分位範囲による閾値に基づき、機械的に得られた数値データを A, B, C の３段階評価に変換する。さらに、総合評価は予測式による計算を行い、A, B+, B, B-, C の５段階評価に変換するという流れになる。

次に、Communicability タスクのライティング・サンプルに関しては、Accuracy タスクと同様に平均文長と難語割合が算出される。アイディア数は、書かれている行数のカウントによる。定型表現 (to+動詞) の頻度と含まれる単語数については、定型表現リストに掲載されている to+動詞の出現数による。観点別評価は数値データを A, B, C の３段階評価に、総合評価は予測式による計算を行い、A, B+, B, B-, C の５段階評価に変換して結果の出力を行うという処理系統になる。

さらに、それぞれのタスクの評価を予測する回帰式の有用性については、回帰式により算出される予測得点をサンプル提供者である 20 名の学生の Criterion スコアと照合することにより検証を行った。それぞれのタスクの回帰式による予測得点と Criterion スコアとの相関係数を計算し、併存的妥当

性の検証を行った結果、Accuracy 評価の 回帰式による予測得点は r = .77, Communicability 評価の回帰式による予測得点は r = .64 となり、いずれも中程度以上の相関が見られた。この結果から、回帰式による予測得点には、ある程度の有用性があると判定された。また、回帰式による予測得点と Criterion スコアとの残差を求めたところ、残差が±1.0 以上となったサンプルは、Accuracy 評価の回帰式による予測得点では 1, Communicability 評価の回帰式による予測得点では 4 サンプルが該当した。このサンプル数からも、自動評価採点システムによる最終的な総合評価を算定する各タスクの予測式には一定の妥当性があると判断することができる。

今後の研究において、本システムの予測精度をより高めるためには、次の (1)～(4) を当面の課題として改良に取り組む必要があると考えられる。

(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、各評価指標の数値データによる予測式の有用性の方が高いことが示唆されたが、サンプル数を増やし実証すること。

(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を考案すること。

(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を考案すること。

本研究により明らかになった(1)～(4)の課題解決に取り組み、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引き続き開発していくことが望まれる。

＜引用文献＞

①Elliot, S. (2003). *How does IntelliMetric score essay responses? RB-929,* Newton, PA: Vantage Learning.

②Kukich, K. (2000). Beyond automated essay scoring, the debate on automated essay grading. *IEEE Intelligent Systems, 15*, 22–27.

③Landauer, T. K., Laham, D. and Foltz, P. W. (2003). Automated scoring and annotation of essays with the intelligent essay assessor, Shermis, M. & Burstein, J (eds.) *Automated essay scoring: A crossdisciplinary perspective,* 87–112, Hillsdale, NJ: Lawrence Erlbaum Associates.

④ 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*, 77–103.

⑤Sugita, Y. (2009b). Developing and improving rating scales for a task-based writing performance test. *JLTA Journal, 12*, 85–103.

⑥Sugita, Y. (2010). Reliability and validity of a task-based writing performance assessment for Japanese learners of English. *JLTA Journal, 13*, 21–40.

⑦ Sugita, Y. (2012a). Effects of rater training on raters' severity, consistence, and biased Interactions in a task-based writing assessment. *JLTA Journal, 15,* 61-80.

⑧ Sugita, Y. (2013). *Comparability of accuracy and communicability tasks: Are they all equally difficult? JLTA Journal, 16*, 67-86.

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 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, 『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』(統計数理研究所共同研究リポート 215) 15-28.

⑪ 杉浦正利 (2008). 英文ライティング能力の評価に寄与する言語的特徴について, 成田真澄(編) 『学習者コーパスに基づく英語ライティング能力の評価法に関する研究』平成 17～平成 19 年度科学研究費補助金(基盤研究 (C) 研究成果報告書) 35-58.

⑫ 杉田由仁 (2012b). 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１件）

① Yoshihito Sugita. "Examination of objective rating indices to an automated essay scoring system for task-based writing tests," 『全国英語教育学 会紀要*(ARELE)*』, Vol. 27, 17-32, 2016.

〔学会発表〕（計３件）

① Yoshihito Sugita.平成 28 年 4 月 14 日 The 50th IATEFL Annual Conference (Birmingham), *Forum on testing* " Objective rating indices to automated essay-scoring systems for writing assessment"

② 杉田由仁 平成 27 年 8 月 23 日 第 40 回全国英語教育学会熊本研究大会 「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における客観的評価指標の検討」

③ Yoshihito Sugita.平成 26 年 8 月 18 日 The 19th Conference of PAAL (Tokyo), "English teachers' perceptions of on-line rater training programs for a task-based writing performance

test.

〔図書〕（計１件）
① 杉田由仁 平成 27 年 2 月 (分担執筆)「第５章　ライティングの評価」『英語教育の実践的探究』149-179, 渓水社

〔その他〕
明治学院大学文学部　英語教授法研究室
http://www.meijigakuin.ac.jp/~ysugita/homepage/

６．研究組織
(1)研究代表者
杉田　由仁（SUGITA, Yoshihito）
明治学院大学・文学部・准教授
研究者番号：70363885