外国語教育メディア学会 (LET) 関西支部 メソドロジー研究部会 2013 年度報告論集
石井 雄隆・近藤悠介 (pp. 1–11)

英語学習者を対象とした自動採点システム―課題と展望―

石井 雄隆
早稲田大学大学院生

近藤 悠介
早稲田大学

概要

本稿では，英語学習者を対象とした自動採点システムについて論じる。学習者の発話や作文の自動採点は評定者間の評価のずれや採点の手間といった評価の際に生じる問題の軽減，あるいは解決に非常に有用であるため，近年盛んに研究が進められている分野である。はじめに，これまでに開発された英語学習者のための発話と作文の自動採点システムについての概略を示し，自動採点の問題点について検討する。

Keywords: 自動採点, SpeechRater, E-rater

1. はじめに

学習者の発話や作文を自動で採点するための研究は，作文においては，Page (1966)，発話においては Bernstein, Cohen, Murveit, Rtischev, and Weintraub (1990)が始まりであると言われている。それ以降，文書処理，インターネット，自然言語処理の三つの影響で作文自動評価の研究は飛躍的に進歩した (Shermis, Burstein, & Bursky, 2013)。近年では，CALICO Journal の 2009 年の On the Automatic Analysis of Learner Language，Language Testing の 2010 年の Automated Scoring and Feedback Systems for Language Assessment and Learning や Assessing Writing の 2013 年の Assessing writing with automated scoring systems などの特集が組まれている。また国内においては，小林・金丸（2012）が作文，小林・阿部（2013）が発話を対象に，それぞれ機械学習に基づく英語学習者の習熟度の自動推定を試みている。また英語だけでなく，日本語の発話の自動採点も今井（2013）などによって開発されている。

自動採点と関連するものとして英語学習者の文法的誤りの自動検出に関する研究が，自然言語処理の分野で盛んにおこなわれている。小町（2013）によると，英語学習者支援のための共通タスクとして，Helping Our Own (HOO) というワークショップが 2011 年と 2012 年に行われており，2011 年は Association for Computational Linguistics (ACL)

Anthology Reference Corpus を用いて，論文の文法的誤り訂正のコンテストが行われた。また 2012 年は，Cambridge Learner Corpus を用いて前置詞と限定詞の文法的誤り訂正を対象として行われた。また Seventeenth Conference on Computational Natural Language Learning (CoNLL 2013)では，限定詞，前置詞，数，動詞の形，一致，スペル，句読点の文法的誤り訂正が The National University of Singapore (NUS) Corpus of Learner English を用いて行われた。学習者の発話，作文の自動採点に関する研究は，外国語教育学，自然言語処理で注目を集めている分野であり，外国語教育学の文脈では，学習者の習熟度を捉えるための有益な示唆が得られる研究である。本稿では，発話と作文の自動採点システムの概略を述べ，問題点について考察する。

2. 発話自動採点システム

2.1 発話自動評価システムの開発動機

外国語の発話の評価はその重要性が認識されているのにも関わらず，いくつかの理由から直接評価する機会が十分に用意されていない。発話自動採点システムの開発に関わる研究者からは次の理由が挙げられている。(1)発話を直接評価するテストを実施するためにかかる人的，時間的が膨大である。(2)評定者による評価のぶれが大きい。(3)評定者の訓練に要する人的，時間的が膨大である (Clauser, Margolis, Clyman, & Ross, 1997; Zechner, Higgins, Xi, & Williamson, 2009)。これらに加え，訓練を行っても評定者による評価のぶれがなくならないことや信頼性の高い評定者を継続的に確保する難しさも学習者の発話が直接評価されていない理由と言える。自動採点システムの実装はこれらの問題を解決し，学習者の発話を直接評価する機会を与えるものである。さらに，評価を伴う発音訓練は膨大な時間が必要であり，コンピュータを用いた自学自習が適しているとの指摘もある (Franco, Bratt, Rossier, Gadde, Shriberg, Abrash, & Precoda, 2010; Witt & Young, 1997)。また，発話自動採点システムの開発は，習熟度の違いによる発話の特徴量の変化や発話誘出タスクが学習者の発話に与える影響を精緻に測定しようとする研究であるため，関連分野に有用な情報を与える。

2.2 発話自動採点システム開発の変遷

作文自動採点が開発された約 20 年後，Bernstein, Cohen, Murveit, Rtischev, and Weintraub (1990)により最初の発話自動採点システムが発表された。この研究では次のような手順により学習者の発話を自動採点した。(1)日本人英語学習者による読み上げ発話を収集する。(2)発話を専門家が評価する。(3)音声認識機を用いて英語母語話者の発話をアライメントする。(4)文節音ごとに学習者と母語話者の発話との差をスコアとして算出する。この自動採点システムが算出したスコアと評定者によるスコアの相関は.80 以上であり，この時点で自動採点システムの実用性が示されたと言える。しかし，大和（2011）

が指摘しているように，英語学習者の発音は英語母語話者をモデルとしないという考えは古くからある。また，英語教師の約 80%を非母語話者が占めている (Canagarajah, 1999)という現状を鑑みると，採点の参照元として英語母語話者の発音を使用することは必ずしも適切であるとは言えない。また，英語母語話者の発音も地域や社会的階級により異なるため，どのような母語話者をモデルにするかという問題もある。

発話の超分節音的特徴を対象とした研究 (e.g. Cucchiarini, Strik, & Boves, 2000a; 2000b)では，学習者の発話の特徴量と評定者による評価を検証し，そこから予測方法を得るという方法が採用されている。話す速さ，ポーズの数などを予測変数，評定者による評価を基準変数として重回帰分析を行い，ここで得た予測式を用いて受検者の評価の予測を行っている。この方法を採用することにより，母語話者を発話のモデルとした場合の問題がある程度解決された。このような研究における評価の予測とは，言い換えれば，評価の高い学習者をモデルにし，このモデルとの近さを判定するという仕組みである。

発話自動採点の採点方法を評定者による評価の予測と考えると，この問題には機械学習の手法を用いることができる。近年の英語教育の趨勢を鑑みると，テストの結果を間隔尺度ではなく順序尺度で与えようとする傾向が見られる。これはヨーロッパ言語共通参照枠 (Common European Framework of References: Council of Europe, 2001)に代表される言語スキルに関する can-do の普及が大きな理由である。このようなことから，発話自動採点システムにおいても順序尺度による評価が算出されることが望まれる。この前提を踏まえると，発話自動採点の予測方法は，いくつかの発話の特徴量をもとにその発話の評価（カテゴリ）を予測するという機械学習の手法が応用できる。実際に，Higgins, Xi, Zechner, and Williamson (2011)では採点可能な発話と不可能な発話の弁別にサポート・ベクター・マシンが利用されている。

3. 英作文自動採点システム

3.1 英作文自動採点システムの概要

本節では，これまでに開発された英作文自動採点システムを概観する。発話の自動採点と同様，作文の自動採点も評定者間の評価のずれや採点の手間といった作文評価の際に生じる問題の軽減，あるいは解決に非常に有用であるため，近年盛んに研究がされている分野である。

英作文の自動採点に関する研究は Page (1966)までさかのぼることができる。このシステムは Project Essay Grade (PEG)と呼ばれ，大規模テストの作文評価における教師 の負担を軽減することが目的であった。PEG の最初のバージョンは，proxes とよばれる約 30 の特徴量を用いており，これらの特徴量を trins とよばれる測定しようとする作文能力の指標の代わりに用いた。

2013 年 4 月現在，少なくとも AutoScore, LightSIDE, Bookette, e-rater, Lexile Writing

Analyzer, Project Essay Grade, Intelligent Essay Assessor (IEA), CRASE, IntelliMetric, BETSY の 10 個の自動採点システムがある (Elliot & Klobucar, 2013; Rudner & Liang, 2002)。本稿では，その中でも e-rater に焦点をあてる。その他のシステムの概要については，石岡（2004; 2009）や Shermis and Burstein (2013)などを参照されたい。

3.2 E-rater

E-rater は，アメリカの Graduate Management Admissions Test (GMAT)の作文試験で用いられており，自動採点システムの中では最も有名なものであると言える。Educational Testing Service (ETS)の Jill Burstein の研究グループが開発をはじめ，2000 年より ETS Technologies に開発と運用が移管されている。専門家と e-rater の評価の一致率は Burstein et al. (1998)では，89%であったが，Burstein and Wolska (2003)では 97%であり，精度はかなり向上している。

E-rater は，Ver.1 では，約 60 の特徴量に基づいて，重回帰分析による採点を行っていた。その際に，全ての特徴量を用いるのではなく，論題によって 10 個程度の変数を選んでいた。2004 年に開発された Ver. 2 は，Burstein, Chodorow and Leacock (2004)によると，以下の 12 個の特徴量を用いて，作文を評価する。

1. 総語数に対する文法エラーの割合
2. 総語数に対する語の使用法についてのエラーの割合
3. 総語数に対する手順のエラーの割合
4. 総語数に対するスタイルについてのエラーの割合
5. 必要とされる談話要素の数
6. 談話要素における平均語数
7. 作文を 6 点法で採点する際に語彙の類似度が一番近い点数
8. 最高点を取った作文との語彙の類似度
9. Type-Token Ratio
10. 語彙の困難度
11. 平均単語長
12. 総語数

この 12 個の変数は，論題によらず，固定されている。E-rater は，これらを基に重回帰分析を行い，学習者の作文を自動で評価している。また e-rater は Criterion というウェブベースの作文支援ツールに組み込まれている。これは，学習者の作文をその場で採点し，診断的なフィードバックを与えることができる。

3.3 E-rater以外の作文自動採点システム

E-rater 以外の自動採点システムは用いている変数や手法が異なる。E-rater のように，重回帰分析を用いているのが，先述した PEG である。また Intelligent Essay Assessor (IEA) は，潜在意味解析，IntelliMetric は決定木分析に基づくルール発見，BETSY はベイズの定理を応用した自動採点を行っており，システムによってさまざまなアプローチが存在する。また，IEA は，Pearson Knowledge Technologies (PKT)として WriteToLearn や Versant という学習システムに組み込まれており，IntelliMetric は Vantage Learning の My Access!という自動作文評価システムに組み込まれている。それ以外にも，無料で使える作文自動添削ツールが大澤 (2013)で紹介されている。

4. 自動採点の問題点

4.1 発話自動採点システムの問題点とその改善方法

自動採点で使用されるタスクおよび評価の観点が音声言語処理の技術に大きく制限されていることが，発話自動採点システムの大きな問題のひとつとして挙げられる (Xi, 2010. p. 294)。音声言語処理の技術では，一般的に発話の自由度が上がるにつれてその認識精度は落ちる。この問題の解決方法として Versant では，テキストで提示した文を読み上げたり，提示された語の反対語を言うタスクを用い，発話の自由度を制限している。Versant は，その測定概念 (construct)を”Facility in L2”とし，Facility は「日常的な事柄について話されている言葉を理解する能力および分かりやすい言葉を使って母語話者同士の会話と同じペースで適切に話すことができる能力」と定義されている(Bernstein, Van Moere, & Cheng, 2010, p. 358)。Versant は，測定しようとする能力を間接的に測定していることとなる。

これに対して，ETS が開発する SpeechRater は，直接測定するものとして自然発話を測定の対象としているが，ここで問題となるのは認識の精度である。SpeechRater の単語認識率は，受検者によって 10%から 80%と大きく異なるが，全体として約 50%と報告されている (Zechner, et al., 2009, p. 888; Xi, et al., 2012, p. 379)。話された単語の 10%しか認識することができない状態で，その発話の単語数やポーズの長さをもとにした変数を用いた評価が妥当な評価であると言うことはできない。

これらの問題を改善するひとつの方法は，認識率の向上である。SpeechRater では隠れマルコフモデルを用いて，母語話者の音声をモデルとした音声認識機を使用し，話者適応を行っていない (Zechner, et al, 2009, p.888-891)。Franco, Abrash, Precoda, Bratt, Rao, Butsberger, Rossier and Cesari (2000)が示すように，モデル学習の際に非母語話者の音声を使用することが認識率向上につながると言われている。また，話者適応の技術を利用することによっても認識率の向上が見られる可能性がある。さらに，近年では，従来の混合ガウス分布を出力確率とした隠れマルコフモデルではなく，深層学習のニューラル・ネット

ワークが算出した出力を使用した隠れマルコフモデルの方が認識の精度が高いことが分かっている (Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath, & Kingsbury, 2012)。これらの技術を用いることでかなりの程度の認識率の向上が期待できる。

もうひとつの改善方法は，複数のタスクを利用して受験者の能力を推定することである。比較的簡単に実装できることから文の読み上げがタスクとして使用されているが，文の読み上げ発話では，受験者によって発話される単語が同じであるため，語彙に関する特徴量や発話量に関しては測定できないが，文節音，韻律などの特徴量を変数として測定することが可能である。また，ある程度発せられる単語が制限される談話完成タスクなどを用いることも可能である。これらのタスクでは測定できない能力を自然発話で測定するというように，いくつかのタスクを組み合わせて，測定したい能力が測定できるテスト・セットを作ることにより，より妥当性，信頼性の高い発話自動採点システムが構築される。タスクの設定により学習者の発話は影響を受けることが知られており，発話を誘出するタスクの作成には応用言語学，第二言語習得，英語教育学の知見が必要である (Kondo & Ishii, 2014)。

4.2 作文自動採点システムの問題点

作文自動採点の問題点は，様々な研究で検討されている (Attali, 2013; Bennett, 2006; Bennett & Bejar, 1998; Clauser, Kane, & Swanson, 2002; Yang, Buckendahl, Juszkiewicz, & Bhola, 2002)。具体例としてまず第一に挙げられるのは，人間の評価との関係性である。自動採点システムの一番の動機づけになるのは，先述した採点の手間の軽減であるため，人間の評価にシステムの評価が近ければ近いほど良いというのは，納得ができる。しかし，これに対して，石岡（2004）はシステムの性能を評価する唯一の基準として人間の評定に必要以上に頼らないことが今後のシステムの要件として望まれると言及している。その理由として，人間の評価は評価基準表に基づいており，それが論題ごとに変わる PEG や IntelliMetric，e-rater (Ver. 1)はモデルとして奇妙であると述べている。E-rater Ver. 2.0 は，論題に関係なく変数は固定されているので，この問題は解決される。一方で，作文の論題が異なっていたとしても，評価に影響を与えない (Brossell & Hoetker Ash, 1984; Spaan, 1993)という立場だけでなく，論題は作文における言語使用に影響を与える (Hinkel, 2009)という立場があることを考えると，論題の影響に対しては，今後さらなる研究が望まれる。

自動採点システムのフィードバックの妥当性について Xi (2010)は，10 個の観点から議論している。ここでは最近の研究が焦点を当てている「自動フィードバックは正確に学習者の産出の特徴や誤りを検出しているか」と「自動フィードバックは学習者にとって有益か」という二つの問題の妥当性を Weigle (2013)に基づきながら検討する。

Weigle (2013, p. 48)は，「自動フィードバックは正確に学習者の産出の特徴や誤りを検

出しているか」という問題に対し，冠詞や前置詞の文法的誤りに関しては精度が上がっていると述べている。しかし，自動採点システムは，全体の意味は明らかだが，評定者の間で意見が割れるような誤りの際には，信頼性をもって誤りを分類できないのではないかと言われている。一例として挙げられているのは，”he lead a good life”という英文である。これが，主語と動詞の一致の誤りなのか，あるいは時制の誤りなのかは文脈によって変わってしまう。またもう一つの問題点として挙げられるのは，評定者は必ずしもある英文上の表現が誤りか誤りでないかで意見が一致するとは限らないということである，特に前置詞の誤りに関してこれは当てはまる。この点に関しては，自動採点システムの構築の際に，何が誤検出で，何が検出漏れになってしまうかを決める必要がある。すなわち，正しいものを正しいと評価でき，間違っているものを間違っていると評価できる度合いを高めるだけでなく，誤りでないものを誤りと判断することを避け，誤りを誤りと判断できない度合いを低めなければならない。これらは学習者コーパス研究の成果を援用，あるいは自動採点研究の成果を学習者コーパス研究に還元するべき点である。

この問題に対して，Criterion の開発グループである Chodorow, Gamon and Terteault (2010)は， Criterion は，80%の精度で前置詞誤りを検出し，冠詞の誤りに関しては，90%の精度を検出したと報告している。Ferris (2006)は，教師と訓練された評定者の誤りの同定率は 80–90%であったと報告しており，この実験が全ての誤りを対象としていたことを考えると，弁別の難しい前置詞や冠詞の誤りで 80-90%の検出ができる Criterion の精度は高いと判断することができる。

同様に，Weigle (2013, p. 49)は，「自動フィードバックは学習者にとって有益か」という問題について下記の二つの文献を引用しながら検討している。Chen and Cheng (2008)は，My Access!という Vantage Learning の自動作文評価システムの教室における使用について検討した。My Access!はスコアとフィードバックを返すが，そのフィードバックに対して学習者は，曖昧，あるいは抽象的であると感じたと報告している。同じように，Grimes and Warschauer (2010)は，自動フィードバックの実施には教師のサポートが不可欠であると述べている。しかし一方で，田地野他（2012）は，Criterion を講義で用いた結果，語彙や文法の正確性，総語数と総文数，論理展開の明確な文章構成といった項目の修正を促し，英文の質的な向上が観察されたと報告している。

5. おわりに

本稿では，発話と作文の自動採点研究の概要と問題点について検討した。自動採点研究は，統計手法の洗練や機械学習の発展に伴い，今後も手法，変数選択，データの組み合わせによって様々な知見が蓄積されていく分野である。一方で，作文の自動採点研究は，作文の専門家からの批判にさらされることがある (Ericsson & Haswell, 2006)。しかし，採点に関わる変数を探求することは教育実践に大きく貢献するのではないだろうか。例えば，

決定木の出力結果などは，そのまま Hirai and Koizumi (2008)で検討されている Empirically derived, Binary-choice, Boundary-definition (EBB)尺度を検討する際に活用することで，人間が評価の際に重きを置く特徴量と，機械が評価の際に重きを置く特徴量の違いを明らかにできる可能性もあるのではないだろうか。そういった教室での貢献の可能性も含めて，自動採点研究は今後も進められていく必要がある。

参考文献

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 181–198). New York: Routledge.

Bennett, R. E. (2006). Moving the field forward: Some thoughts on validity and automated scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 403–412). Mahwah, NJ: Erlbaum.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–17. doi:10.1111/j.1745-3992.1998.tb00631.x

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. *Proceedings of the International Conference on Spoken Language Processing*. 1185-1188.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377. doi:10.1177/0265532210364404.

Brossell, G., & Hoetker Ash, B. (1984). An experiment with the wording of essay topics. *College Composition and Communication, 35*(4), 423-425.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In B, Christian., & Whitelock, P. (Eds.), *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp.206–210). Burlington, MA: Morgan Kaufmann Publishers.

Burstein, J., & Wolska, M. (2003). Toward evaluation of writing style: finding overly repetitious word use in student writing. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics,* Budapest, Hungary, 35–42. doi:10.3115/1067807.1067814

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine, 25*(3), 27–36. doi:10.1609/aimag.v25i3.1774

Canagarajah, A. S. (1999). Interrogating the "native speaker fallacy": non-linguistic roots, non-pedagogical results in G. Braine (Ed.). *Non-native educators in English language teaching*.

Mahwah, NJ: Lawrence Erlbaum Associates.

Chen, C.-F. E. C., & Cheng, W.-Y. E. C. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94–112.

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of grammatical error detection systems for English language learners: Feedback and assessment. *Language Testing, 27*(3): 419–436. doi:10.1177/0265532210364391

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*(4), 413–432. doi:10.1207/S15324818AME1504_05

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement, 34*(2), 141–161.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: CUP.

Cucchiarini, C., Strik, H., & Boves, L. (2000a). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication, 30*, 109–119. doi:10.1016/S0167-6393(99)00040-0

Cucchiarini, C., Strik, H., & Boves, L. (2000b). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of Acoustical Society of America, 107*(2). 989–999.

Elliot, N., & Klobucar, A. (2013). Automated Essay Evaluation and the Teaching of Writing. In Shermis, M., & Burstein, J. (Eds.) *Handbook of automated essay evaluation* (pp. 16–35). New York: Routledge.

Ericsson, P. F., & Haswell, R. (Eds.) (2006). *Machine scoring of student essays: Truth and consequences.* Logan, UT: Utah State University Press.

Ferris, D. R. (2006). Does error feedback help students writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge, UK: Cambridge University Press.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., ... & Cesari, F. (2000). The SRI EduSpeak™ system: Recognition and pronunciation scoring for language learning. *Proceedings of InSTILL 2000*, 123–128.

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided

language learning applications. *Language Testing, 27(3)*, 401–418. doi:10.1177/0265532210364408

Grimes, D. & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 1-44.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language, 25*(2), 282–306. 10.1016/j.csl.2010.06.001

Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics, 41,* 667-683. doi:10.1016/j.pragma.2008.09.029

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97.

Hirai, A., & Koizumi, R. (2008). Validation of an EBB scale: A case of the Story Retelling Speaking Test. *Japan Language Testing Association Journal, 11,* 1–20.

石岡恒憲 (2004)「記述式テストにおける自動採点システムの最新動向」『行動計量学』, 31(2), 67–87.

石岡恒憲 (2009)「論述式項目の自動採点」植野 真臣・永岡慶三 (編)『e テスティング』(pp. 95–120). 培風館.

今井新悟 (2013)『音声認識技術を応用したコンピュータ自動採点日本語スピーキングテストの開発』科学研究費助成事業（科学研究費補助金）研究成果報告書.

小林雄一郎・阿部真理子 (2013). 「スピーキングの自動評価に向けた言語項目の策定」『電子情報通信学会技術研究報告』 113(253), 1–6.

小林雄一郎・金丸敏幸 (2012). 「Coh-Metrix とパターン認識を用いた課題英作文の自動評価」 『人文科学とコンピュータシンポジウム論文集―つながるデジタルアーカイブ』 259–266.

小町守 (2013)「ウェブマイニングを用いた英語学習支援」人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会発表資料

Kondo, Y., & Ishii, Y. (2014). Bridging the Gap Between Second Language Acquisition Research and the Development of Automated Scoring System for Second Language Speech. In R. C-H. Tsai. & R. Guy (Eds.), *Language, Culture, and Information Technology* (pp. 149–164). Taipei, Taiwan: Bookman Books.

大澤真也 (2013)「学習者の自律的な推敲を促す自動添削ツールの検討」外国語教育メディア学会関西支部 2013 年度秋季研究大会発表資料

Page, E. B. (1966). The Imminence of Grading Essays by Computer, *Phi Delta Kappan, 47*, 238–243.

Rudner, L. M. & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1*(2), 1–12.

Shermis, M. & Burstein, J. (2013). (Eds.) *Handbook of automated essay evaluation.* New York: Routledge.

Shermis, M., Burstein, J. & Bursky, S. (2013). Introduction to Automated Essay Evaluation. In Shermis, M., & Burstein, J. (Eds.) *Handbook of automated essay evaluation* (pp. 1–15). New York: Routledge.

Spaan, M. (1993). The effect of prompt in essay examinations. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 98-122). Alexandria: TESOL.

田地野彰・細越響子・川西慧・日髙佑郁・髙橋幸・金丸敏幸 (2012). 「アカデミックライティング授業におけるフィードバックの研究−Criterion®を導入した授業実践からの示唆−」『京都大学高等教育研究』, 17, 97–108.

Weigle, S. (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.

Witt, S. & Young, S. (1998). Computer-aided pronunciation teaching based on automatic speech recognition. In S. Jager, J.A. Nerbonne, & A.J. van Essen (Eds.), *Language teaching and language technology,* 25–35. Lisse: Swets & Zeitlinger.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291–300. doi:10.1177/0265532210364643.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing, 29(3)*, 371–394. doi:10.1177/0265532211425673.

大和知史. (2011). 「L2 speech 研究における発音の「明瞭性」の取り扱い―明瞭な評定のために―」『外国語教育メディア学会中部支部メソドロジー研究部会 2011 年度報告論集』. 41–49.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15,* 391–412. doi:10.1207/S15324818AME1504_04

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication 51*, 883–895. doi:10.1016/j.specom.2009.04.009.