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　５ 年　 ６ 月 １５ 日現在

機関番号：１７２０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ００７４２

研究課題名（和文）CBT対応の英語発話自動採点システムの構築：指導と評価一体化の高大接続に向けて

研究課題名（英文）Automated Speech Scoring of Dialogue Response by Japanese Learners of English as a Foreign Language

研究代表者
林　裕子（Hayashi, Yuko）

佐賀大学・教育学部・准教授

研究者番号：１０６４９１５６

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では、「話す」能力の直接評価に向けた取組として自動採点システムの導入に着眼し、談話完成タスク（Discourse Completion Task：DCT)を用いて発話自動採点システムの構築と運用に取り組んだ。DCTをCBT（Computer-Based Testing）形式で大学生210名に実施した。人手と自動評価の一致率を基に算定した本システムの予測精度は72%であり、教室環境などのローステークス環境で運用できる可能性が示唆された。加えて、受験から採点結果提供までの全プロセスを自動化したシステムを用いて実証実験を実施した。その結果を基に教室環境での幅広い運用について提案する。

研究成果の学術的意義や社会的意義
今日、音声認識技術に加えて表情やモーションキャプチャーの技術も搭載し、バーチャル・エージェント（virtual agent）と英語で学習を進める双方向（やり取り）型の学習支援システムの開発・運用が進んでいる。しかし、（「学習」ではなく）自動「採点」（＝「話す」能力の測定）において、継続・発展性のある文脈が用いられている例は皆無に近い。本研究では、今日の外国語教育で重視されている、文脈や目的に応じた情報の理解と伝達が求められる談話完成タスクを用いた自動採点のシステム構築・精査に取り組み、外国語教育における自動評価の運用の可能性について検討・提案を行える点で学術的・社会的意義に富むと言える。

研究成果の概要（英文）：Numerous theoretical and/or practical problems make it challenging for teachers, researchers and policymakers to assess speaking proficiency in contexts that reflect real conversational situations. Using automated speech scoring technologies is one way to alleviate these problems. This study set out to build and evaluate a new system for automatically assessing learners' speech within the context of an oral Discourse Completion Task (DCT). The DCT was administered to 210 undergraduate students of intermediate English language proficiency. The results show that the exact agreement between human and machine scores was moderately high, 72 %. This value is comparable to the extant literature on automated speech scoring and could serve as a foundation upon which to explore the applicability of the system in classroom settings. Areas for improvement in prediction accuracy and possible ways of untilising the current system in pedagogical settings are discussed.

研究分野：外国語教育（言語テスティング、自動採点研究）

キーワード：発話自動採点　スピーキング　外国語教育　談話完成タスク

１．研究開始当初の背景

外国語教育では、小・中・高等学校を通じ、現実社会と密接に関連づいた真正な文脈における概念理解の深まりや知識・スキルの総合的な活用を評価するパフォーマンス評価の実践が進められている。その一方で、4 技能評価を実施している大学は少なく、指導と評価が一体化した高大接続には課題が多く残されている。特に「話す」能力の評価については、信頼性の保証や実施に伴う膨大な時間と労力の面から、継続的な実施は容易ではない（石井・近藤，2020）。これらの課題の打開策の一つに自動採点システムの導入が挙げられる。

発話自動採点システムでは、自然言語処理や機械学習の技術を用いて、学習者の発話から得られた言語的特徴量（総語数、異語数、文長など）を基に、習熟度のレベルや、テストの得点を推定する。SpeechRater（Zechner, et al., 2009）や Ordinate (Pearson Education, 2018)の既存のシステムでは、一般的に、音声認識（大量のデータで構築された音響モデルの音声的特徴と学習者の音声的特徴との照合）において計測された発話特徴量から評定者が与える点数を予測する手法が用いられる。これらのシステムでは、発話の自由度が低いタスク（例 : read aloud, sentence repetition）と高いタスク（例 : stating opinions, retelling a lecture）の双方が用いられているが、①単発的な（一問一答型の）文脈が大部分を占める、②自由発話では、学習者が発する単語の認識率が低下するなどの課題点が指摘されている（Evanini, et al., 2015）。

近藤・石井（2017）はこれら（①&②）の問題の克服に向け、大学生を対象に以下に示すような談話完成タスク（Discourse Completion Task: DCT）を用いて発話自動採点システムを構築した。

❖ You (A) want to end your conversation. What would you say in the conversation below?
A: _____.　（例：　‘It was nice talking to you.’, ‘Well, I have to get going now.’）
B: See you.

評定者は、1（適切）及び 0（不適切）の２値で発話を採点した。その結果、自動評価と人手評価の一致率は 74%であり、既存のシステム（SpeechRater）を用いた例 (Zechner et al., 2009）より高い数値が得られた。同システムの単語認識率は 71%と高い数値が得られ、教室環境などのローステークス（low-stakes）文脈での活用に向けて有望な結果が示されている。

２．研究の目的

近藤・石井（2017）が開発したシステムでは、発話の自由度が制限された DCT が使用されているため先述した①の課題克服には至っていない。実際の言語使用をより正確に反映させるためには、継続・発展性のある文脈を用いたタスクが必要である。今日、音声認識技術のほかに、表情やモーションキャプチャーの技術も搭載し、バーチャル・エージェント（virtual agent）と英語で学習を進める双方向（やり取り）型の学習支援システムの開発・運用も進んでいる（例：*AutoTutor,* Graesser, 2016)。しかし、‘（「学習」ではなく）自動「採点」（＝「話す」能力の測定）において、継続・発展性のある文脈が用いられている例は非常に少ないため。そこで、本研究で、次の２つを目的とする。１つ目は、文脈や目的に応じた情報の理解と伝達が求められる口頭による Discourse Completion Task (DCT)を CBT(コンピュータ上で行なう試験：Computer-Based Testing) の形式で実施し、その発話データを用いて発話自動採点システムを開発することである。２つ目は、同システムの予測精度の検証結果から、英語コミュニケーション能力の評価・測定におけるタスクや評価基準の精査と設定を行い、英語教育への導入の可能性とその方法について検討・提案することである。

３．研究の方法

3.1 研究課題

本研究では、近藤・石井（2017）を基に以下の研究課題を設定し、DCT を用いた発話自動採点システムの構築に取り組んだ。

1. DCT を用いて測定する英語学習者の「話す」能力は、 発話自動採点システムによってどの程度予測することができるか。
2. 本システムは教室環境において、どの程度運用することができるか。

3.2 受験者

教養教育や専門教育の一環で英語の授業を履修する合計 210 名（男性 108 名、女性 102 名）の大学生が参加した。内、28 名は授業の重複から複数回録音を行ったため、合計 238 個の音声データファイルを収集した。母集団を代表するサンプルとなるよう、複数の学部を対象に実施した。

3.3 Discourse Completion Task

本研究では、近藤・石井（2017）を参考に、談話の連続性を高めるよう 3 往復（3 つのターン）で構成される DCT（所要時間約 10 分）を作成した。作成した DCT はタブレット端末で実施できるよう、アプリケーション化した。使用した DCT（音声データ）を以下に示す。

❖ あなた(C)は友人２人(A&B) と旅行の計画を立てています。相手の発話を聞き、会話の流れに合うように答えなさい。

A: We're graduating soon. It's sad because we'll be in different cities.
B: I know. Hey, we should go somewhere together to make memories.
A: Sounds good! Do you have any suggestions for where to go?
C: Q1. _____ 【Turn 1】
A: Do you know how to get there?
C: Q2. _____ 【Turn 2】
A: When and How long is this holiday going to be?
C: Q3. _____ 【Turn 1】

3.4 手続き

授業担当者の協力の下、研究者と補助員 2 名の 3 名で計 7 クラスを訪問し、説明、機器の配付・回収を含め、授業時間の 30 分以内で実施した。収集した音声データ（714 発話＝238 発話×3 ターン）は文字起こしを行い、日本語話者 2 名と英語話者 2 名の合計 4 名で、発話内容の明瞭さと適切さについて、高校生と大学生を対象とした予備調査（早瀬他, 2017）をもとに開発した表 1 のルーブリックを用いて 0〜3 の 4 値で採点を行った。評定者は二手（Pair 1:日本語話者 1 名、英語話者 2 名，Pair 2:日本語話者 1 名、英語話者 1 名）に分かれ、それぞれ、357 発話（119 発話 ×3 ターン）の採点を行い、高い採点者間信頼係数が確認された（Kappa (κ) = .91 (p < .001), 95% CI (confidence intervals) [.87, .94] (Pair 1) and κ = .91 (p < .001) [.86, .92] (Pair 2)）.

４．研究成果

4.1 【研究課題１】自動採点システムの予測精度について

評価基準、評価の方針、評定者による評価値が付与された発話サンプルを参考にし、評定者と協議の上、評価を予測する特徴量（予測変数）を決定した（表２）。高得点者は発話総語数（token）や発話文の複雑さ（complexity）の値も高いことや、言い淀み(hesitation)が多い発話者は評価が低いなどの傾向を踏まえ、(1)発話の総語数、 (2)um”、“uh” などの言い淀み(hesitation)の総数、(3) 複雑さ（complexity：節および句の数）、(4) 音響尤度（confidence score：本研究で使用した音響モデル（Watson IBM Speech-to-Text）と発話音声の類似度）の 4 つを予測変数として採用した。

上記の予測変数を用いて、多項ロジスティック回帰を実施し評価を予測した。発話データの 80%を学習データ、残りの 20%をテストデータとし、学習されたモデルを用いたテストデータにおける予測精度は 72%であった（図１）。同採点モデルにおける各変数の貢献度を図 2 に示す。

表 1. DCT ルーブリック

3	■	対話の８割以上において、その目的や場面に応じた内容で明瞭かつ適切に応答している。
	■	対話相手の質問に対し、具体的に Good news（良い知らせ）（Q1）、その知らせを受け遠く離れてしまうことへの気持ち（Q2）を具体的に述べることができている。
2	■	対話の 5 割以上において、その目的や場面に応じた内容で適切に応答している。
	■	対話相手の質問に対し、具体的に Good news（良い知らせ）（Q1）、その知らせを受け遠く離れてしまうことへの気持ち（Q2）について、前者のみ適切に述べることができている。または、Q1 と Q2 も概ね適切に応答しているが、明瞭さに課題が見られる。
1	■	対話の 3 割において応答しているが，明瞭さや内容の適切さの面で課題が見られる。
	■	1 つ目の質問には応答できているが、2 つ目の応答が不明瞭、不適切であり、課題が顕著にみられる。
0	■	無回答、明瞭さや内容の適切さが判断できない。
	■	1 つ目の質問に対し、沈黙を貫いたり、不明瞭、不適切な応答で答えており、やり取りが断絶している。

表 2. 発話データの特徴量

	得点	総語数（token）			言い淀み	音響尤度	複雑さ（complexity）		
		ターン（T）					ターン（T）		
N=238	(0-3)	T1	T2	T3	（hesitation）	(confidence score)	T1	T2	T3
平均値	2.2	20.59	17.21	17.96	3.29	0.56	11.03	7.68	7.44
標準偏差	1	12.74	11.86	11.6	10.45	0.15	6.18	6.27	5.54
最小値	0	0	0	0	0	0	0	0	0
25%	2	11	8	9	0	0.47	7.14	0	3.23
50%	3	20	17	17	0	0.57	11.11	7.41	7.69
75%	3	29	24.5	26	2.94	0.66	14.91	12.13	10.81
最大値	3	69	64	64	100	0.85	33.33	25	25

図 1. 人手評価と自動評価の一致度

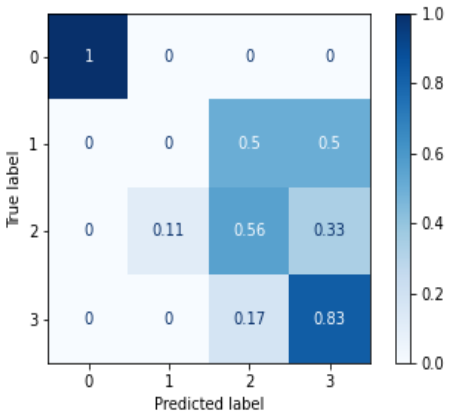
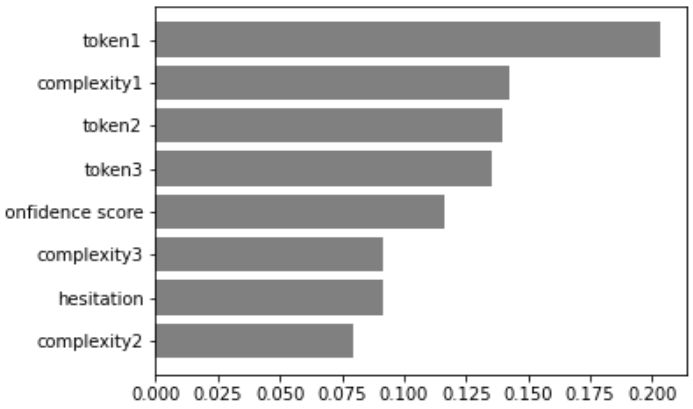


図 2. ターン別特徴量の重要度



*ラベルの数値＝ターン (Turns 1-3).

4.2【研究課題２】教室環境での運用について

4.1 で検証したシステムの予測精度は 72%と、教室環境や練習などのローステークス文脈で運用するには十分な数値が得られたことから、(4.1 では採点結果の提供部分は自動化されていない) 受験から採点結果の提供まで全てのプロセスを自動化してアプリケーション化し、4.1 と同様にタブレット端末を用いて CBT 形式で実施した。対象者は 4.1 の中から Focus Group として 17 名を選定し、通常の授業時間（合計 90 分）の 30 分を利用して実施した。全自動化した本システムの概要を図 3 に示す。受験者は DCT を受験した数分後に「結果閲覧システム」にログインし、即時的に結果（得点＋記述文）を閲覧することができる。

受験後に本システムの使用感や有用性についてアンケートを実施し質的に分析を行った結果（図 4)、概ね好意的な回答が得られているが、自由記述欄において実施環境（「他の受験者の声が気になる（n＝5)」）やルーブリック（「他の検定試験との対照表がほしい（n＝3)」、「100 満点（n＝2）などもっと細かい基準で数値化してほしい」などの回答が得られた。

５．総合考察とまとめ

本研究の結果から、連続性のある談話においても、自由度が制限された DCT を用いた場合（近藤・石井，2017）と同等の予測精度を有する自動採点システムの構築が可能であることが示された。しかし、本研究で用いたシステムには、改良の余地が残されていることも明らかになった。図 1 が示すように、高得点の受験者については人手評価と自動評価の一致度は高い一方で、得点 1 については一致度が 0%である。これには、そもそも得点 1 の発話が少なったこと（全体の 10%）のほかほか、ルーブリックの精度が関係していることが要因として考えられる。得点 1 の記述文は「対話の 3 割において応答しているが，明瞭さや内容の適切さの面で課題が見られる」と設定しているが（表 1)、最初のターンで対話が成立し（適切な発話が提供され）なければ以降のターンにおける発話内容は評価対象とならないことから、2, 3 番目のターンにおける発話内容が十分に反映されていない可能性がある。実際に、本研究の採点モデルで用いた予測変数の中でも最初のターン (Turn 1) における総語数や複雑さの寄与率が高いことが示されている (図 2)。得点が低い（得点 1， 0）受験者は共通して、‘uh’, ‘erm’などの言い淀みが多く見られた。言い淀みは流暢さのみならず、明瞭性にも影響する発話特徴である (Chen et al., 2018; Jenkins, 2020)。

図 3. 自動採点システム概要

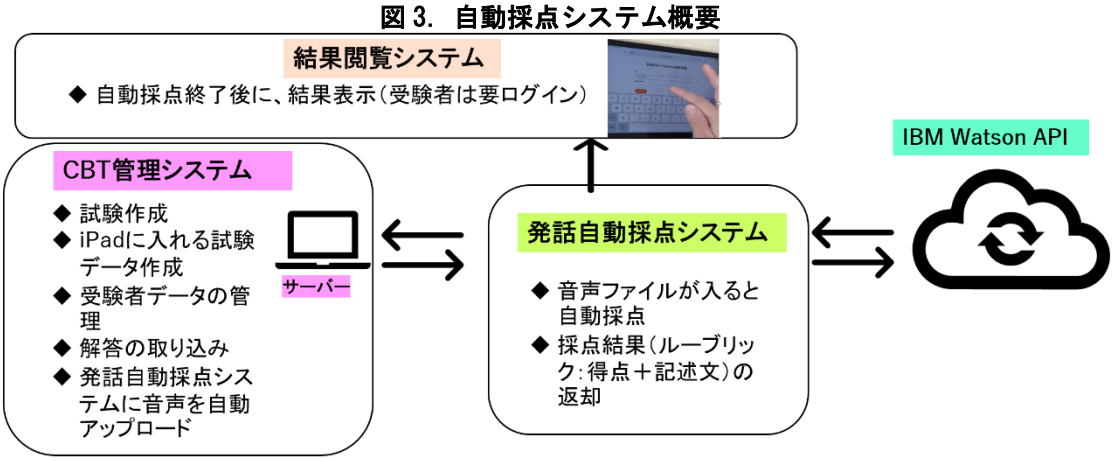
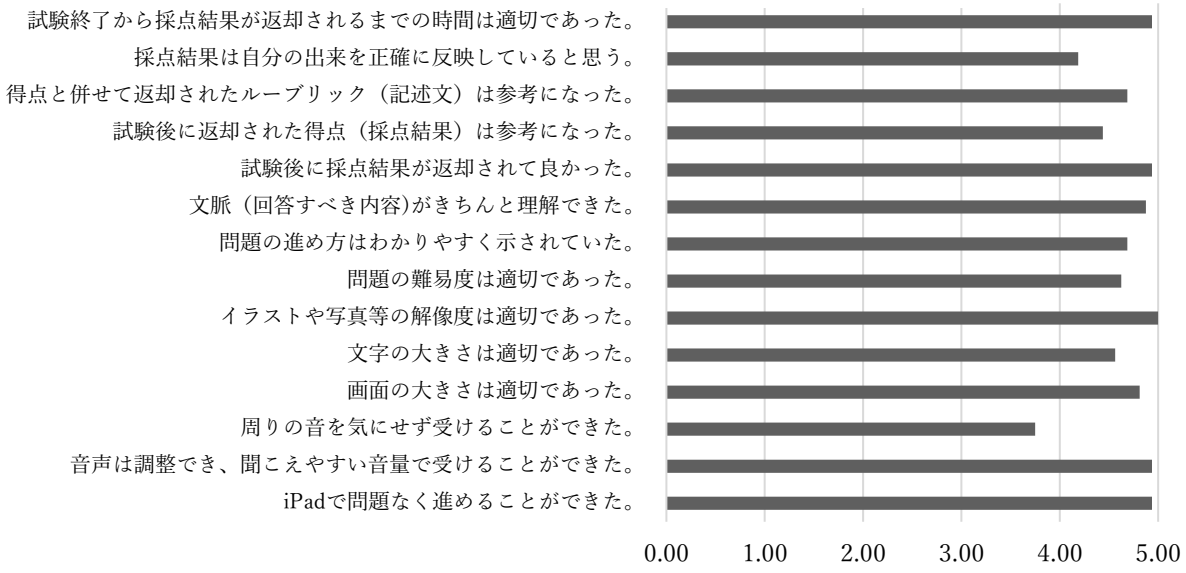


図 4. 事後アンケートの結果



1 全くそう思わない、2 あまりそう思わない、3 どちらともいえない、4 そう思う、5 非常にそう思う

本研究では、受験者の発話の明瞭性は音響尤度（confidence score）のみで測定したが、先行研究では、発音の質（母音や子音の持続時間）や単語の強勢位置などの音律特性も明瞭性に影響することが示されている（Zoghbor, 2018; Chen et al., 2018）。特に、強勢に関しては、日本人英語学習者を対象とした先行研究において、語句単位に限らず文強勢（リズム）の指標も考慮する必要性が指摘されている（Tepperman et al., 2010）ことから、より明瞭性についての詳細な分析が今後の課題となる。

少人数での実施ではあるが、①授業時間内に実施できた（約 30 分）、②ほぼ即時的（数分以内）にフィードバック（採点結果）の提供が行えた、③学習者のニーズとの一致度についての質的分析が行えた点において、教室環境における本システムの導入に向けた有益な情報が得られたと言える。しかし、本研究は 1 種類の DCT を用いたシステム検証のみに留まってしまっているため、今後タスクの数と種類を増やし、タスク間におけるシステムの精度及び採点モデルの汎用性の検証やルーブリックの改良に取り組む必要がある。

【引用文献】

Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., & Ma, M. (2018). Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine. *ETS Research Report Series*, *2018*(1), 1-31. https://doi.org/https://doi.org/10.1002/ets2.12198

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). Automated scoring for the TOEFL Junior® comprehensive writing and speaking test. *ETS Research Report Series*, *2015*(1), 1-11. https://doi.org/https://doi.org/10.1002/ets2.12052

Graesser, A. C. (2016). Conversations with AutoTutor Help Students Learn. *International Journal of Artificial Intelligence in Education*, *26*(1), 124-132. https://doi.org/10.1007/s40593-015-0086-4

早瀬 博範・林 裕子・ 江口 誠（2018）「四技能を問う英語 CBT 入試開発に向けた取組み」『LET Kyushu-Okinawa Bulletin』, 18, 15-29. https://doi.org/10.24716/letko.18.0_15

石井雄隆・近藤悠介（編）(2020). 『英語教育における自動採点 現状と課題』ひつじ書房.

Jenkins, J. (2020). Where are we with ELF and language testing? An opinion piece. *ELT Journal*, *74*(4), 473-479. https://doi.org/https://doi.org/10.1093/elt/ccaa045

近藤悠介・石井雄隆 (2017).「英語学習者の発話自動採点システムの開発と英語教育プログラムへの導入可能性の検討」『Language Education & Technology』54, 23-40.

Pearson Education. (2018). *Pearson Academic PTE scoring guide*. Retrieved March 10 from https://laxmioverseas.com/wp-content/uploads/2019/09/score-guide.pdf

Tepperman, J., Stanley, T., Hacioglu, K., & Pellom, B. (2010). *Testing suprasegmental English through parroting*. Paper presented at Speech Prosody, Chicago, IL.

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*(10), 883-895. https://doi.org/10.1016/j.specom.2009.04.009

Zoghbor, W. S. (2018). Teaching English pronunciation to multi-dialect first language learners: The revival of the Lingua Franca Core (LFC). *System*, *78*, 1-14. https://doi.org/https://doi.org/10.1016/j.system.2018.06.008

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　3件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Yuko Hayashi, Yusuke Kondo & Yutaka Ishii	４．巻 (Published online)
２．論文標題 Automated speech scoring of dialogue response by Japanese learners of English as a foreign language	５．発行年 2023年
３．雑誌名 Innovation in Language Learning and Teaching	６．最初と最後の頁 1-16
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/17501229.2023.2217181	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 林　裕子	４．巻 56
２．論文標題 英語発話自動採点システムの精度と教室環境での使用について	５．発行年 2023年
３．雑誌名 Highway the Bulletin of English Teachers in Saga	６．最初と最後の頁 2-3
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 無
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 林　裕子	４．巻 55
２．論文標題 英語発話自動採点システムの構築と実用性について	５．発行年 2022年
３．雑誌名 Highway the Bulletin of English Teachers in Saga	６．最初と最後の頁 3-4
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 無
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 林　裕子	４．巻 54
２．論文標題 小・中・高を通じた英語習得に向けて	５．発行年 2020年
３．雑誌名 Highway the Bulletin of English Teachers in Saga	６．最初と最後の頁 3-4
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 無
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 早瀬博範、林裕子、江口誠	４．巻 18
２．論文標題 四技能を問う英語CBT 入試開発に向けた取組み	５．発行年 2018年
３．雑誌名 LET Kyushu-Okinawa BULLETIN	６．最初と最後の頁 15-29
掲載論文のDOI（デジタルオブジェクト識別子） 10.24716/letko.18.0_15	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　0件）

１．発表者名 林　裕子
２．発表標題 談話完成タスクを用いた英語発話自動採点システムの構築と運用
３．学会等名 第５０回九州英語教育学会　佐賀研究大会 KASELE SAGA
４．発表年 2022年

１．発表者名 林　裕子・近藤　悠介・石井　雄隆
２．発表標題 談話完成タスクを用いた英語発話自動採点システムの構築
３．学会等名 全国英語教育学会第46回長野研究大会
４．発表年 2021年

１．発表者名 林　裕子・近藤　悠介・石井　雄隆
２．発表標題 談話完成タスクを用いた英語発話自動採点システムの構築と実用化
３．学会等名 SONAS 2022 Symposium on Automated Scoring
４．発表年 2022年

〔図書〕　計3件

1．著者名 石井雄隆・近藤悠介	4．発行年 2020年
2．出版社 ひつじ書房	5．総ページ数 156
3．書名 英語教育における自動採点 現状と課題	

1．著者名 林　裕子・竜田　徹	4．発行年 2021年
2．出版社 東京書籍	5．総ページ数 120
3．書名 よくわかる！教師を目指すための高大接続のしくみ	

〔産業財産権〕

〔その他〕

―

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	近藤　悠介 (Kondo Yusuke) (80409739)	早稲田大学・グローバルエデュケーションセンター・准教授 (32689)	
研究分担者	石井　雄隆 (Ishii Yutaka) (90756545)	千葉大学・教育学部・准教授 (12501)	

	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
研究協力者	村上　明 (Murakami Akira)		

6．研究組織（つづき）

	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
研究協力者	モクスン　ジョナサン (Moxon Jonathan)		
研究協力者	ルウ　ペトラス (Roux Petrus)		

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
英国	University of Birmingham			