特集：ライティング評価の新潮流

ライティング評価のための自動評価研究の展望と課題

李　在鎬

要　旨

本稿では、ライティング評価に関する自動評価研究の課題と展望を明らかにするため、次の４点を述べた。１）自動評価の研究はパフォーマンス評価の課題を解決するために出現したこと、２）自然言語処理の技術を利用しているため、自動評価の結果は、誤解析があることを前提に理解すべきこと、３）自動評価に用いられる習熟度を予測するための計算モデルは近似値であること、４）自動評価に用いられる分析指標は固定的な定量的なものが使用されていること。そして、この４点の特徴を踏まえ、自動評価研究のさらなる活性化のため、１）大規模なデータ構築が必要であること、２）データ分析のリテラシーを持つ日本語教育の実践者を養成する必要があること、３）自動評価のための多様な分析指標を発掘する必要があることを述べた。

キーワード

自動評価　自然言語処理　信頼性　データ科学　学習者コーパス

１．自動評価とは何か

自動評価とは、「話す」や「書く」のように産出に関わる言語活動をコンピュータの計算モデルで処理する技術の総称である。その起源とされる研究はPage（1966）の研究であり、英語教育を中心に約50年にわたる研究の歴史がある（石岡 2007、近藤・石井（編）2020）。日本語教育分野では、システム構築に関する研究（李ほか 2017、田中ほか 2017、李ほか 2019）、システムの妥当性に関する研究（小森ほか 2018、小森ほか 2022）、応用や実践に関する研究（村田・李 2018、影山 2019）などがあるが、いずれもここ数年に出た論文ばかりであり、英語教育に比べ、研究の量・質のいずれにおいても十分ではなく、これから研究の活性化が期待される分野と言える。なお、文献によっては、自動採点（Automated Scoring）と称されることもある。

以下では、自動評価の位置づけを明らかにするため、次の４点について述べる。１点目はパフォーマンス評価との関係、２点目は自然言語処理との関係、３点目は採点に使用さ

れる計算モデルの課題、４点目は自動評価の分析指標に関する問題である。

1.1　パフォーマンス評価と自動評価

「聞く」や「読む」のような理解に関わる言語活動の場合、多肢選択式の客観テストによって、その言語能力の程度を判断することができる（李（編）2015）。一方、「話す」や「書く」のように、産出に関わる言語活動の場合は、直接、話させたり書かせたりして、その能力を評価する方法が用いられる[1]。

「話す」テストとしては、「OPI（Oral Proficiency Interview；鎌田ほか 2020）」が広く用いられている。「書く」テストに関しては、「OPI」のような汎用的なテストはないが、課題遂行を目的とする作文タスクを設定し、構成を意識しながら書かせるタスクをし、その成果物を評価することが多い（関・平高（編）2013、田中・阿部 2014、望月ほか 2015）。こうした評価は、現実に起こりうる言語使用の課題がどの程度、遂行できるかを測る目的でなされるものであり、「パフォーマンス評価」と総称されている（近藤 2012）。

パフォーマンス評価は、言語技能を直接的に観察できる点から真正な評価方法として認識されている（Hart 2012、李（編）2015）。このパフォーマンス評価は真正な評価であるというメリットがある反面、採点が難しいというデメリットがある。というのは、客観テストのように答えを一つにしぼることができないため、一貫性のある採点が難しいからである。一般的には訓練された評価者（採点者）が回答を目視で確認し、主観に基づいて評価を行う。主観に基づいて評価を行うことはパフォーマンス評価の課題として認識されている。なぜなら、主観は統制が難しく、定量的でないため、主観評価には何らかの不安定さが伴うからである。こうした課題が存在するため、パフォーマンス評価は、大規模なテストや受験生の人生を左右するハイステークテスト（High Stake Test：ブラウン 1999）で実施する場合に、膨大なコストがかかる。

石井・近藤（編）（2020: 4）は、主観評価に基づくパフォーマンス評価の問題として、以下の５点を指摘している。

１．試験の実施、評価に人的、時間的コストがかかる。
２．評価の信頼性を保つことが困難である。
３．評価者の訓練に人的、時間的コストがかかる。
４．訓練した評定者を長期に確保することが困難である。
５．「話す」能力を評価する評定者には高い能力が必要である。

こうした課題を解決するものとして、自動評価が提案された。アプリケーション化された自動評価システムでは、人のかわりにコンピュータを使っているため、常に一貫性のある動作をする。また、コンピュータ上で動作するため、24時間休むことなく動き続けることができ、運用コストも少なくて済む。さらに、大規模なデータを高速に処理できるのもコンピュータの重要な特性である。このように人のかわりにコンピュータを使うことで、信頼性のある評価をより少ないコストで実現できる、というのが自動評価の研究を推進する合理的な理由になっている。

1.2　自然言語処理と自動評価

パフォーマンス評価の課題を解決するために出現した自動評価の研究は、自然言語処理（Natural Language Processing: NLP）と計量言語学の研究成果を基盤に成り立っている。特に、自動評価の結果を理解するためには、自然言語処理の技術が持つ特徴について理解する必要がある。

自然言語処理とは、人の言語をコンピュータ上で処理するために立ち上がった研究領域で、1940年代後半からスタートした分野である。形態素解析、構文解析、意味解析、文生成、言い換え、機械学習、知識獲得などの要素技術を使い、機械翻訳（自動翻訳）、情報検索、文書処理、情報抽出、音声・マルチモーダル情報処理のような応用システムを開発・提案している（言語処理学会 2009）。日本語の自然言語処理について言えば、90年代に、「ChaSen（茶筌）[2]」に代表される形態素解析[3]の技術が確立し、2000年代には、ウェブの普及により、情報検索に関わる技術が確立した。そして、2010年代には、機械翻訳などの応用システムが実社会で活用されるようになり、現代の情報化社会を支えている。

こうした自然言語処理の技術を利用した学習支援システムは、2000年代以降、様々なものが開発されている（李 2021）。自動評価に関するシステムも、その延長線上にあると捉えることができ、自然言語処理の要素技術なしでは、成立しない。例えば、日本語学習者作文評価システム*jWriter*（https://jreadability.net/jwriter/、Lee & Hasebe 2020）の場合、文章解析の最初のステップで、形態素解析を行い、平均的な１文の長さを計算したり、漢語や和語の使用率を計算したりして、それをもとに文章の習熟度（proficiency）を推定するのである。

自然言語処理の解析は、コンピュータを利用して行うため、大量のデータを高速で処理できるメリットがある反面、解析結果には、一定程度の誤解析が含まれている点を忘れてはならない。そして、学習者の文章には誤用が含まれているため、形態素解析の誤解析の程度はさらに大きいことを認識する必要がある（李 2009）。こうした理由から、自動評価の結果を利用する際には、常に解析上のエラーが含まれていることを認識する必要がある。

1.3　計算モデルと自動評価

パフォーマンス評価の課題を解決するため、自然言語処理の成果を利用して行われる自動評価の研究では、人間が予め用意した訓練用のデータをもとに何らかの計算モデルを作るという方法が用いられる（石井・近藤（編）2020）。この計算モデルを用いて、新規文章に対して何らかの予測を行うのである。こうした予測のための計算モデルは、できるだけ現実の事象と乖離がないように、調整されている。パフォーマンス評価の文脈でいえば、各々の計算モデルは熟練の評価者と同程度の評価ができるよう、調整されているのである。とはいうものの計算モデルの結果は、あくまで熟練の評価者の近似値にすぎないという限界がある。

近似に関する具体例を示す。ETS（Educational Testing Service、https://www.ets.org/）が開発したe-Rater®（https://www.ets.org/erater/）という自動評価のための解析エンジンがある。e-Rater® は入力文の文法的誤りや総語数や単語の平均的な長さなど12の素性[4]を抽

出し、重回帰分析に基づいて作成した計算モデルで自動評価を行っている（Burstein et al. 2004: 32）。この計算モデルの予測精度を評価するため、専門家の評価とe-Rater®の評価の一致度を調べたところ、97％であり、専門家同士が主観評価の一致度に匹敵するものである[5]。

1.4　自動評価の分析指標

e-Rater®のような自動動評価エンジンは、近似ではあるものの、（作文タスクによっては）限りなく専門家の評価に近い判断をするが、人とコンピュータでは評価のプロセスが本質的に異なることを理解しなければならない。具体例を示す。以下の文章１と文章２に注目してほしい。

文章１）

ここ十数年の間に日本の雑誌の発行部数はどんどん少なくなっています。また新聞販売契約数も、特に若年層において低下し続けています。これらの背景には、インターネット技術の急速な発展・普及があり、この傾向はますます強まると考えられています。

確かに今、あらゆるニュースはインターネット上で即座に、かつ安易に入手することが出来ます。例えば日本最大手であるヤフーの場合、そのトップページに載ったニュースが新聞で取り上げられるのは翌日になってからだと言われています。つまりヤフーは新聞が報じるより一日早くニュースを利用者に提供できる訳です。

また、インターネットの利点として、瞬時に様々な角度からの意見を閲覧出来ることが挙げられます。インターネット上では、利用者なら誰でも何らかの形で、自ら情報を発信することが可能です。そのため、一つのニュースをとっても、賛成論・反対論、積極論・消極論のように、あらゆる立場の人間の意見を目にすることが出来る訳です。雑誌・新聞では、何誌・何紙も集めなければ、このような事は困難です。

しかし、雑誌・新聞にもインターネットにない長所があります。インターネットに比べて匿名率が低いこと（つまり誰がその記事に責任を持つのかが判りやすい）、情報格差を生じさせにくいこと（インターネットを使えない、特に高齢層に対応できる）などです。また、サーバーのダウンなどのシステムトラブルに左右されないという点も重要だと言えます。

このようにインターネットと、雑誌・新聞はそれぞれ、私達の社会に欠くことが出来ない長所を持っていると言えます。そのため、いかにインターネットがニュース源として定着・発展しようと、雑誌・新聞は必要とされ続けるでしょうし、無くなってしまうというとこは有り得ないと、私は考えます。

文章２）

今、「インターネットでニュースを見ることができるから、もう新聞や雑誌はいらない。」という意見を持っている人がたくさんいるけど、私はそうは思いません。

確かに、インターネットで多くの資料をさがることができるし、すぐ世界中の情報を知っているし、とても便利です。けれども、インターネットは人に「実感」を与えられません。人間にとって、「実感」はとても重要な感覚で、人は「実感」がある物事の中に安全感や真*実を感じられます。でも、人はインターネットで何も触れないし、何か持っている気持ちを感じられません。そのゆえに、新聞や雑誌などの書物が必要です。

人にとって、新聞や雑誌はただ情報を集める物だけではなく、新聞や雑誌は人の生活の一部分になりました。もし新聞や雑誌がなくなったら、人はどこかおかしくなった気持ちを感じる可能性もあります。

こうして、新聞や雑誌は単純な工具*ではなく、人の習慣や感情の一つです。また、新聞や雑

誌には収蔵性がある。ある新聞や雑誌は単純な書物ではなく、芸術性もあります。紙の質から版面のデザインまではちゃんと考えられた物です。もしインターネットで読んだら、それらの芸術性はもたいなくなります。
それから、人はただ視覚の生物ではないです。人にとって、確かにインターネットは便利ですけど、インターネットにも限制性があります。新聞や雑誌の紙のにおいと触感はまたとないです。
だからこそ、世界のあらゆるの物はまたとなくの特質があります。新聞や雑誌もそうです。もう新聞や雑誌はいらないわけにはありません。

上記の二つの文章は、伊集院ほか（2020）の分析で用いられた作文データで、「インターネットが自由に使える時代に新聞や雑誌は必要か」[6]というテーマで書かれた「意見文」である。伊集院ほか（2020）では、これらの文章に対して40名の専門家による主観評価の結果を報告しているが、それによると文章１は上位群、文章２は下位群として評価された。この結果を踏まえ、*jWriter*で自動評価を行ってみた。その結果、文章１は「上級」（評価値：3.47）、文章２は「中級」（評価値：2.11）と判定され、文章２よりは文章１が良いという結果になっている。なお、*jWriter*では「1.637＋平均文長×0.045＋中級後半語×0.021＋タイプトークン比×－0.430＋動詞×0.015＋中級前半語×0.011＋総文字数×－0.004＋和語×0.007＋漢語×0.007」（Lee&Hasebe 2020）という計算式に基づいて評価値を計算しているが、評価値が高ければ高いほど、文章としての習熟度も高いことになる。

さて、人の主観評価とコンピュータの計算モデルは、結果としては一致しているが、結果にたどり着くまでのプロセスは大きく異なる。特に評価に利用するリソースの面において大きく異なる。人の主観評価では、文章としての結束性、主張の明示さ、根拠の有無、内容としての面白さといった要素が重視される（伊集院ほか 2020）。一方、*jWriter*のような自動評価の場合、平均文長や文字数や動詞や語種の使用頻度などの要素に基づいて評価をする。つまり、コンピュータが用いる評価指標は、固定的で定量的な性質を持つが、人が用いる評価指標は、動的で定性的な性質を持つのである。

２．主観評価と自動評価

人間が行う主観評価とコンピュータが行う自動評価の類似点および相違点は、表１のようにまとめることができる。

表１　人間とコンピュータの評価

	人間による主観評価	コンピュータによる自動評価
誰が	熟練の評価者が	コンピュータ上の計算モデルが
何を	学習者のパフォーマンスを	学習者のパフォーマンスを
なぜ	言語学習を支援するため	言語学習を支援するため
どうやって	評価者の主観に基づいて	自然言語処理の解析結果と計量言語学の分析指標に基づいて
どこで	対面空間で	仮想空間で

表１は、人間による主観評価とコンピュータによる自動評価を対比させて整理したものである。両者の違いは、一見すると人間が行ってきたことをコンピュータにやらせているだけのように見える。しかし、研究の基盤レベルで異なることも注目しなければならない。特に研究分野と研究成果の活用の面で考えてみたい。

まず、研究分野面で見た場合、主観評価の研究では、応用言語学を主軸にして、教育学、心理学、統計学にまたがる形で研究が行われる。一方、自動評価の研究の場合、データ科学を主軸とし、情報学、教育工学にまたがる形で研究が行われる。こうした分野の違いから研究の基本となる視座の違いが生じる。というのは、主観評価の研究では、教育活動との連関という視座が重視されるのに対して、自動評価の研究では、どのような訓練データで、どのような計算モデルを構築したのか、そして、予測の精度はどの程度かということが重視されるからである。

次に、研究成果の活用の面で見た場合、主観評価は、対面空間が前提になるため、研究成果の活用も局所的であることが多いが、自動評価の場合、研究成果はウェブシステムをとおし、不特定多数の学習者に対して発信される。つまり、研究成果の影響を受ける範囲の面において、主観評価は限定的であるのに対して、自動評価は広範囲であるとみることができる。

３．自動評価研究の活性化のために

自動評価に関する研究や普及を促すため、３つのアプローチが必要である。１つ目は、大規模なデータの構築、２つ目は、データ分析のリテラシーを持つ日本語教育の実践者の養成、３つ目は、自動評価のための多様な分析指標の発掘である。

１つ目のアプローチとして、自動評価の研究を推進するためには大規模な学習者の産出データが不可欠である。いわゆる学習者の産出データを大規模に集めた学習者コーパスが必要である。日本語教育の分野でいえば、「I-JAS（International Corpus of Japanese as a Second Language）」（迫田ほか 2020）のような例が挙げられる。「I-JAS」のようなこれまでの学習者コーパスは、何らかの文字列検索をして、用例を目視して、研究をするという想定で作られているため、限られた規模のものであった。しかし、自動評価のようにコンピュータ上で情報を読み込み、機械学習の方法で計算モデルを作るとなると、現状のものより、数倍は大きい訓練データが必要となる。（少し乱暴な表現が許されるならば）自動評価の研究を進めるための学習者コーパスの開発においては、質より量を重視したコーパス開発が必要と言える。

２つ目のアプローチとして、自動評価を活用できる日本語教育の実践者が必要である。具体的には、推測統計や多変量解析について基本的なアルゴリズムを理解し、自然言語処理の技術的課題についても理解を持つ日本語教育実践者が必要である。こうした人材が自動評価の可能性と課題について理解を持ち、自らの教育実践の場で主体的に自動評価を活用する必要がある。そして、教育実践の成果が実践研究として共有されていくことが重要と言える。

３つ目のアプローチとして、自動評価のための新たな分析指標の発掘が必要である。自

動評価で用いられる指標の多くは、計量言語学の成果を踏まえたもので、形態素や語彙の単位で指標化がなされてきた。しかし、言語運用を捉えるためには、より長い単位に基づく指標や文を超える単位の分析指標が必要である。

4．最後に：自動評価との向き合い方

人工知能との共生が叫ばれる現在の社会状況を考えた場合、自動評価は、言語教師にとってもっとも身近な存在の一つである。自動評価は、（主観評価の負担軽減を目的に提案されたためか）便利さだけが強調されがちであるが、利用する立場においては、自動評価のメリットとデメリットの両方を把握した上で使うことが重要である。そして、自動評価の取り入れ方に関しても、主観評価と並行した使い方が良いと考える。

自動評価と主観評価の使い分けとしては、次のような方法が考えられる。診断的評価や総括的評価のように明示的な得点化が必要な場合は、自動評価を取り入れたほうが良く、形成的評価のように指導目的で評価を行う場合は、主観評価を取り入れたほうが良い。その理由としては、次のことが考えられる。診断的評価や総括的評価の場合、信頼性が重要であり、このことに関しては、コンピュータの計算モデルは優れていると言える。一方、教師による主観評価ではルーブリックを活用しながら、より良い文章を書くためのサポートをすることが重要で、これは教師にしかできない仕事である。

このような視点に基づいて自動評価と主観評価の在り方を考えた場合、自動評価は、できるだけシンプルで単層的な形で結果を示すほうがよく、主観評価は、できるだけ多層的で豊かな情報を与える形で結果を示すほうが良い。

付記　本研究はJSPS科研費19H01273、19K21637の助成を受けて行われました。

注

1　実際に書かせたり、話させたりする方法（直接テスト）のほかに間接テストの方法もある。「書く」でいえば、実際に作文を書かせる代わりに語句を並び替えさせて記号で解答させる整序作文などがある（望月ほか 2015）。「話す」でいえば「SPOT」（小林 2015）や日本語能力試験の「即時応答」（板橋 2020）などのように即座で反応できるどうかを測定することで、間接的に話す能力を測定することができる。

2　「形態素解析システム『茶筌』version 2.4.0 使用説明書」（https://ja.osdn.net/projects/chasen-legacy/docs/chasen-2.4.0-manual-j.pdf/ja/2/chasen-2.4.0-manual-j.pdf.pdf）、（2022.7閲覧）

3　「形態素（morpheme）は，言語の意味や文法機能を担う最小の単位と定義される．日本語の形態素解析（morphological analysis）とは，与えられた文を形態素の単位に分割し，その文法機能（一般には品詞および活用情報）を同定する処理を言う．」（言語処理学会2009:138）

4　12の素性は、次のとおりである。1）Number of grammar errors ÷ essay length, 2）Number of usage errors ÷ essay length, 3）Number of mechanics errors ÷ essay length, 4）Number of style diagnostics ÷ essay length, 5）Number of required discourse elements, 6）Average length of discourse elements ÷ essay length, 7）Score assigned to essays with similar vocabulary, 8）Similarity of vocabulary to essays with score 6, 9）Number word types ÷ number of word tokens, 10）Log frequency of least common words, 11）Average length of words, 12）Total number of words

5　計算モデルは、データの量と質によって精度が変わる。今後、大規模な学習者データが集積されることとディープラーニング（Kelleher 2019）のような高度な解析アルゴリズムが普及し、より高精度な計算モデルが提案されていくことが期待される。こうなると、人間の能力を超える計算モデルの登場も想定できるが、専門家同士の評価と一致したかどうかをもとに、計算モデルの精度を計算する方法の妥当性についても再考する必要がある。

6　データの詳細は「日本・韓国・台湾の大学生による日本語意見文データベース」（http://www.tufs.ac.jp/ts/personal/ijuin/terms.html、2022.7.30参照）

参考文献

石岡恒憲（2007）「コンピュータによるエッセイ、小論文の自動採点について」『教育テスト研究センター第３回研究会報告書』、pp.1-8.（https://www.cret.or.jp/files/ea9d28ab0986f97bdf33db02f943871f.pdf、2022.8.閲覧）

伊集院郁子・李在鎬・小森和子・野口裕之（2020）「評価コメントに見られる意見文評価の様相―共起ネットワーク及びコレスポンデンス分析に基づく考察―」『第二言語としての日本語の習得研究』23、pp.26-43

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

板橋貴子（2020）「日本語能力試験聴解「即時応答」における解答プロセス」『国際交流基金日本語教育紀要』16、pp.17-28

影山陽子（2019）「作文の自動評価システムの日本人学部大学生への活用可能性―評価への納得度と推敲への動機に着目して―」『アカデミック・ジャパニーズ・ジャーナル』11、アカデミック・ジャパニーズ・グループ研究会、pp.28-36（http://academicjapanese.jp/dl/ajj/ajj11_28-36.pdf、2022.7閲覧）

鎌田修・嶋田和子・三浦謙一（編）（2020）『OPIによる会話能力の評価 ―テスティング、教育、研究に生かす―』凡人社

言語処理学会（編）（2009）『言語処理学事典』共立出版

小泉利恵（2018）『英語４技能テストの選び方と使い方―妥当性の観点から』アルク

小林典子（2015）「SPOT」、李在鎬（編）『日本語教育のための言語テストガイドブック』くろしお出版、pp.110-126

小森和子・李在鎬・長谷部陽一郎・鈴木泰山・伊集院郁子・柳澤絵美（2018）「教師による評価とコンピュータによる自動評価はどの程度一致するのか―中上級日本語学習者の意見文の評価を対象に―」『2018年度日本語教育学会秋季大会予稿集』、pp.278-283

小森和子・伊集院郁子・李在鎬（2022）「日本語学習者の作文における自動評価と教師評価の比較」『明治大学国際日本学研究』14-1、pp.41-67

近藤ブラウン妃美（2012）『日本語教師のための評価入門』くろしお出版

迫田久美子・石川慎一郎・李在鎬（編）（2020）『日本語学習者コーパスI-JAS入門―研究・教育にどう使うか』くろしお出版

関正昭・平高史也（編）（2013）『テストを作る』スリーエーネットワーク

田中真理・阿部新（2014）『GoodWritingへのパスポート』くろしお出版

田中真理・阿部新・影山陽子・佐々木藍子・坪根由香里（2017）「ヨーロッパ日本語学習者のライティング（エッセイ）分析―総合的評価とマルチプルトレイト評価結果を参照して」『第21回ヨーロッパ日本語教育シンポジウム報告・発表論文集』、pp.75-92

村田裕美子・李在鎬（2018）「読解力と作文力の相互関連性に関する統計的分析」（2018ドイツ語圏大学日本語教育研究会シンポジウム発表資料）

望月昭彦・深澤真・印南洋・小泉利恵（2015）『英語４技能評価の理論と実践―CAN-DO・観点別評価から技能統合的活動の評価まで』大修館書店

李在鎬（2009）「タグ付き日本語学習者コーパスの開発」『計量国語学』27-2、pp.60-72

李在鎬（編）（2015）『日本語教育のための言語テストガイドブック』くろしお出版

李在鎬・長谷部陽一郎・迫田久美子（2017）「人工知能の仕組みを利用した学習者作文評価システム『jWriter』―I-JAS を利用した試み」『2017 年度日本語教育学会秋季大会予稿集』、pp.289-291
李在鎬（2021）「書くことを支援する自動評価システム『jWriter』（特集AIやICTが変える言語教育）」『日本語学（2021 冬号）』40-4、pp.42-51
李在鎬・長谷部陽一郎・村田裕美子（2019）「学習者作文の習熟度に関する自動判定とWebシステムの開発について」、李在鎬（編）『ICT×日本語教育―情報通信技術を利用した日本語教育の理論と実践』くろしお出版、pp.38-53
Brown, J. D.（1996）. *Testing in language programs*. New Jersey: Prentice Hall Regents.（和田稔（訳）（1999）『言語テストの基礎知識―正しい問題作成・評価のために』大修館書店）
Burstein, Jill & Chodorow, Martin（2004）Automated Essay Evaluation: The Criterion Online Writing Service, *AI Magazine*, 25（3）, pp.27-36
Hart, Diane（1992）*Authentic Assessment: A Handbook for Educators*, New York: Addison-Wesley Press（田中耕治（訳）（2012）『パフォーマンス評価入門―「真正の評価」論からの提案』ミネルヴァ書房）
Kelleher, John D（2019）*Deep Learning*（*The MIT Press Essential Knowledge series*）, Cambridge, Mass.: MIT Press（柴田千尋・久島聡子（訳）（2020）『ディープラーニング』ニュートンプレス）
Lee, Jae-Ho, & Hasebe, Yoichiro（2020）Quantitative Analysis of JFL Learners' Writing Abilities and the Development of a Computational System to Estimate Writing Proficiency. *Learner Corpus Studies in Asia and the World*, 5, pp.105-120（http://www.lib.kobe-u.ac.jp/repository/81012493.pdf、2022.7閲覧）
Page, B. Ellis（1966）The Imminence of Grading Essays by Computer, *The Phi Delta Kappa*, 47（5）, pp.238-243

（り　じぇほ　早稲田大学日本語教育研究科）