調査報告
影山陽子/アカデミック・ジャパニーズ・ジャーナル 11(2019)28-36

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―

影山陽子

要旨

昨今、日本語教育の分野でも機械による自動評価システムが開発されている。本研究では、外国語としての日本語学習者向けに開発された作文の自動評価システムを、日本人学部大学生を対象に試用し、アンケートを実施、その結果を分析した。研究目的は次の 3 点である。(1) 自動評価システムは日本人学部大学生（日本語母語話者）の書いた意見文に対して、どのような評価を表出するのか。(2) 日本人学部大学生は、その評価にどの程度納得をするのか。(3) 自動評価システムの日本人学部大学生への活用可能性はあるのか。結果、(1) マルチ評価「日本語」やホリスティック評価は高くなり、「目的・内容」「構成・結束性」では能力に適した評価点が表出された。(2) 評価点が高い場合は、納得度も高いが推敲への動機は生まれず、低い場合は評価基準や方法への疑義や推敲への動機が高まる。(3) 自動評価システムの日本人学部大学生への活用では、学生の独力での活用は難しく、教師による支援や指導が必要であることがわかった。

キーワード

意見文、自動評価、日本人学部大学生、評価への納得度、推敲

1. 研究の背景と目的

1.1 機械による作文の自動評価の可能性

2020 年度から行われる「大学入学共通テスト」では、知識・技能だけでなく、大学入学段階で求められる思考力・判断力・表現力を一層重視するという考えに基づき、国語や数学に記述問題が課されることが決定している（独立行政法人大学入試センター 2019)。しかし、その評価方法については、試験自体の規模の大きさや公平性の面から「どのように採点されるのか」等、不安視する声があがっているのが現状である。

一方、英語の大規模試験である TOEFL iBT writing では、評価は採点官による採点に加え、機械による自動採点が活用されている（ETS 2019)。大規模試験における機械による自動評価の利用は、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ことから、近年盛んに研究が進められている分野であり（石井・近藤 2013)、今後、日本語の作文評価においても大きな流れとなっていくことが予想される。

また、作文の教育指導に目を転じてみても、英語教育では自動評価を用いた指導が行われており、学生が自動評価の評価およびフィードバックをどのように受け止めているか等についても研究がなされている（齋藤 2017)。一方、日本語教育においては、学習者作文自動システム J-writer（李他 2017）や教師支援のための日本語ライティングの自動評価システム GoodWriting Rater（田中他 2017）等、作文の自動評価システムの開発が緒に就

いたばかりで、その試用についての研究はこれから行われようとしている段階である。

1.2 研究目的

前節で述べたように、今後日本語作文を対象とした自動評価システムは、採点方法としても指導ツールとしても大きな可能性が見込まれることが予想されるが、日本語作文は日本語学習者が書いた作文と日本語母語話者が書いた作文の 2 つに大別される。しかしながら、自動評価システムの開発に関して、日本語母語話者が書いた作文を対象としたものは、現在、評価モデルの構築過程であり「最終的な評価判断を導き出すことについて扱っていない」(藤田他 2012)。そのため、本研究では、日本語学習者の作文を対象に開発された日本語ライティングの自動評価システム GoodWriting Rater を用いて、日本語母語話者である日本人学部大学生が書いた意見文を対象に評価点データとその受け止めに関するアンケートデータを収集する。そして、それらを以下の 3 つの観点から考察することを目的とする。

(1) GoodWriting Rater は日本人学部大学生（日本語母語話者）の書いた意見文に対して、どのような評価を表出するのか。
(2) 日本人学部大学生は、その評価にどの程度納得するのか。
(3) GoodWriting Rater の日本人学部大学生への活用可能性はあるのか。

2. 日本語ライティングの自動評価システム GoodWriting Rater

2.1 GoodWriting Rater の特徴

GoodWriting Rater は、比較論証型の意見文を対象に自動評価を行うオンラインシステムである。その特徴は、人間によるライティング・パフォーマンス評価と機械学習に基づくライティング・レベル自動推定とを融合させたシステムであること、つまり人間が学習者のライティングを評価し、機械にその結果を学習させるという構築方法を採用した点にある。この人間による評価と機械学習による自動推定の融合は、日本語教育分野では初めての試みであった。

また、そもそもの Good Writing Rater の開発目的は、学習者数が少なく教師が点在している欧州の大学で教える日本語教師が日本語作文指導に苦慮している欧州日本語教師の支援のためであった（田中他 前掲)。そのため、研究の一環として欧州の学習者が書いた日本語作文の収集が行われ、自動評価システム構築の基礎データとして使用されるとともに、教師や学習者が参照できるよう、主に欧州[(1)] 日本語学習者が書いた日本語ライティングのレベル別サンプル提示が HP 上にもなされている（GoodWriting.jp 2019)。いい換えれば、この自動評価システムは学習者個人による使用ではなく、教師のいる授業内での使用を想定し開発されたことも、ひとつの特徴であるといえるだろう。

2.2 GoodWriting Rater の機能

GoodWriting Rater は、サイト上の入力枠に「400 字以上 1600 字以下の日本語」を入力し、「実行」をクリックすると、次の 3 つが表示されるという機能をもっている。

(1) 自動評価の結果
「ホリスティック評価」と「マルチプルトレイト評価」(以下マルチ評価)

(2) テキスト情報

(3) メタ言語情報

まず、(1) 自動評価の結果「ホリスティック評価」と「マルチ評価」に関しては、以下のような説明がされている。

> GoodWriting プロジェクトではホリスティック評価（Holistic scoring）という作文全体の評価と、マルチ評価（Multiple-trait scoring）と呼ばれる観点別の評価スコアを定義しています。マルチ評価では「目的・内容」「構成・結束性」「日本語」の 3 つの観点から評価します（ただし、「目的・内容」は与えられたプロンプトを用いて作文をした場合にのみ有効な観点です)。それぞれ 1-6 の 6 段階でスコアづけします。公開版のシステムでは、低いレベルである 1-2 と高いレベルである 5-6 は区別せず、1-2・3・4・5-6 の 4 段階でスコアを予測します。(GoodWriting.jp 前掲)

次に、(2) テキスト情報では、総文字数、総文数、総段落数、漢字率、ひらがな率、カタカナ率、総文字数÷総文数＜文あたりの平均文字数＞、第 1 段落の文数÷総文数＜全体に対する第 1 段落の割合＞、最終段落の文数÷総文数＜全体に対する最終段落の割合＞の 9 観点が表示される（GoodWriting.jp 前掲)。

さらに、(3) メタ言語ハイライトでは、投稿された作文上に、使われたメタ言語が種類別に色分けでハイライトされ提示される。なお、この場合の「メタ言語」とは「本文の内容とは直接関係のない、文章の展開を理解しやすくするような機能を持つ表現や説明のこと（田中・阿部 2014)」を指している。

自動評価システムにおいては、これらの機能を教師や学習者がどのように使いこなせるのか、その機能を使ってどのようにライティング・パフォーマンスを伸ばせるのかが重要であるが、現在の研究状況はシステム開発とその紹介にとどまっており、授業では使用されていない。そのような状況であるため本実践研究は、1.2「研究の目的」で述べた事情も含め、作文の書き手が本来の対象とは異なるものの、これらの機能を授業内で試用してみることを最優先事項とした。

3. 授業実践とデータ

3.1 対象者とデータ収集方法

今回の授業実践の対象者は、都内単科女子大学の学部 2 年生 122 名である。言語表現科目の 1 コマ（90 分）内、PC 教室にて各自 1 台ずつ PC を用い、意見文執筆→自動評価作業→推敲→再び自動評価作業→PC 入力によるアンケート回答の順序で活動し、データを収集した。自動評価システムの試用に関しては教師がその方法を説明し、かつ、このシステムが外国語としての日本語学習者を対象に開発されたことも説明した。

また、アンケートでは以下のことを聞いた。

①「ホリスティック」は何点でしたか。
②「ホリスティック」の点数への納得度は？
③「目的・内容」は何点でしたか。
④「目的・内容」への納得度は？

⑤「構成・結束性」は何点でしたか。
⑧「構成・結束性」の点数への納得度は？
⑨「日本語」は何点でしたか。
⑩「日本語」の点数への納得度は？
⑪この「自動評価システム」の評価点への感想を聞かせてください。
⑫「メタ言語ハイライト」への印象や感想を聞かせてください。
⑬どんなことに気を付けて「修正」しましたか。
⑭修正後、点数は変わりましたか。
⑮どんな風に変化しましたか。教えてください。
⑯アンケートデータを研究に使用する場合、使用に了解をいただけますか。

対象者 122 名のうち、アンケート未完者 10 名分、研究使用への不承諾者 5 名分を除いた 107 名分の意見文 107 編を本研究の対象データとする。

3.2 プロンプト

プロンプトは、GoodWriting.jp 内に提示されている以下のものを使用した。サイト内に提示されているプロンプト（4 種類）は、自動評価システムの構築のためのデータ収集時に使用されたものであり、適正な評価を得るためにもこれらを用いることが理に適っていると考えられたからである。

外食派と自炊派
あなたは以下の作文コンテストのポスターを見ました。そして、この作文コンテストに応募することにしました。

あなたは「外食派」？それとも「自炊派」？
「外食」と「自炊」、それぞれのプラス面とマイナス面を挙げて比較し、「食生活」についてのあなたの意見を 600 字～800 字で書いてください。
応募者の中から抽選で 20 名様に、弊社のレストラン★★のランチ券（2 名様分）または弊社の自炊グッズ（フライパンと鍋）を差し上げます。
★★食品会社マーケティング部外食派と自炊派

4. 結果と考察

4.1 評価結果

GoodWriting Rater が示すホリスティック評価とマルチ評価（「目的・内容」「構成・結束性」「日本語」）の各評価結果を表 1 に示す。

表 1　「ホリスティック評価」と「マルチ評価」の点数別作文数

	ホリスティック評価	マルチ評価「目的・内容」	マルチ評価「構成・結束性」	マルチ評価「日本語」
5-6 点	77　(72%)	46　(43%)	55　(51.4%)	89　(83.2%)
4 点	15　(14%)	34　(31.8%)	32　(29.9%)	10　(9.3%)
3 点	14　(13.1%)	25　(23.4%)	19　(17.8%)	8　(7.5%)
1-2 点	1　(0.9%)	2　(2%)	1　(0.9%)	0　(0%)
計	107 編（100%）	107 編（100%）	107 編（100%）	107 編（100%）

ホリスティック評価に関しては、77 編（72%）が 5-6 点、15 編（14%）が 4 点と、4 点以上の高評価が 84%であった。

マルチ評価「日本語」は 89 編（83.2%）が 5-6 点、10 編（9.3%）が 4 点と、4 点以上の高評価が 92.5%であり、3 点が 8 編（7.5%）、1-2 点は皆無であった。これは日本語学習者を対象とした自動評価システムを日本語母語話者が使用したからであろう。また、先に述べたホリスティック評価に関しても、日本語能力の高さがホリスティック評価の点数を高めている傾向にあるのではないかと推測される。

一方、マルチ評価「目的・内容」に関しては、5-6 点が 46 編（43%）と半数以下となり、4 点が 34 編（31.8%）、3 点が 25 編（23.4%）、1-2 点が 2 編（2%）となる。マルチ評価「構成・結束性」では、5-6 点が 55 編（51.4%）と約半数となり、4 点が 32 編（29.9%）で、4 点以上が約 8 割であるものの、3 点が 19 編（17.8%）と 2 割弱存在している。

以上から、マルチ評価「日本語」やホリスティック評価においては日本語母語話者であることが有利に働くものの、マルチ評価「目的・内容」や「構成・結束性」においては、ライティング能力の差が評価点の違いとなって表れていることがわかる。

4.2　評価への納得度

次に、学生たちが評価に対してどの程度納得しているかについて考えたい。アンケートでは、自動評価システムが示した各評価の点数（4 段階）を尋ねた後、「○○の点数への納得度は？」という質問をし、5「大変納得できる」から 1「全く納得できない」の 5 段階から 1 回答を選択してもらった。表 2 はその平均値を示したものである。

表 2　各評価点への納得度（平均値）

	ホリスティック評価	マルチ評価「目的・内容」	マルチ評価「構成・結束性」	マルチ評価「日本語」
平均値	4.2	4.0	4.1	4.4

納得度の平均値からは、自動評価の評価点が高かったものほど学生の納得度が高い傾向がみえてくる。

次に「この『自動評価システム』の評価点への感想を聞かせてください」という質問に対する自由記述回答の一部を、各評価の点数（4 段階）と納得度尺度（5 段階）を併記し紹介する。表示する内容は、自由記述回答＋【ホリスティック評価点数（納得度)、マルチ評価「目的・内容」点数（納得度)、マルチ評価「構成・結束性」点数（納得度)、マルチ評価「日本語」点数（納得度)】である。

学習者 A：どうやって評価してるのか気になった。
【ホリ 3 点（4)、目・内 3 点（4)、構・結 5 点（5)、日 5 点（5)】
学習者 B：自分の言葉遣いが的確に点数化されていて、見直さなければいけないところをもう一度見つめ直すことができるため、良いものだと思った。
【ホリ 5 点（5)、目・内 4 点（4)、構・結 5 点（5)、日 5 点（5)】
学習者 C：コンピュータに評価されるのは実際に先生に見てもらうのとはやはり、観点が違ったりするため、均一ではあるが、納得のいくものといかないものがあると思った。
【ホリ 3 点（3)、目・内 3 点（3)、構・結 3 点（3)、日 5 点（4)】
学習者 D：詳しく解析してくれるから良いと思った。
【ホリ 5 点（5)、目・内 5 点（5)、構・結 5 点（5)、日 5 点（5)】
学習者 E：どういうところを見て判断してるのか記述があったらいいと思う。
【ホリ 5 点（5)、目・内 5 点（5)、構・結 4 点（3)、日 5 点（5)】
学習者 F：正しく評価されていると思う。
【ホリ 5 点（4)、目・内 3 点（3)、構・結 3 点（3)、日 5 点（5)】

各評価点の点数と納得度との関係を見ると、納得度の平均値と同様、高い評価点には納得度も高くなっている様子が窺える。自由記述からは、「正しく評価されていると思う」「的確に点数化されている」「詳しく解析していくれるから良い」という感想がある一方で、「どうやって評価しているのか気になった」という疑問や「どういうところを見て判断しているのか記述があったらいいと思う」「コンピュータに評価されるのは実際に先生に見てもらうのとはやはり、観点が違ったりするため、均一ではあるが、納得のいくものといかないものがあると思った」等、評価に対するなんらかのフィードバックを求めるものが見られた。また、このような疑義は評価点が低く表された場合に生まれる傾向があるようだ。

学生たちの評価点への受け止めに関しては、高い評価点が示された場合は納得し、それでよしと受け止めて終わってしまうが、評価点が低かった場合は、評価の方法や基準、その評価点が導き出された理由などを知ろうとする動機が生まれることがわかった。

4.3 推敲の動機と観点

次に、「どんなことに気を付けて『推敲』しましたか」という質問に対する自由記述回答の一部を、前節同様、各評価の点数（4 段階）と納得度尺度（5 段階）を併記し紹介する。

学習者 G：満点だったから修正してない。

【ホリ 5 点（5)、目・内 5 点（5)、構・結 5 点（5)、日 5 点（5)】

学習者 H：点数が低いところを中心に見ていった。

【ホリ 4 点（4)、目・内 3 点（4)、構・結 4 点（4)、日 5 点（5)】

学習者 I：目的、内容のところが 4 だったので、目的に沿って話が進められていたかを修正しようとした。

【ホリ 5 点（5)、目・内 4 点（4)、構・結 5 点（5)、日 5 点（5)】

学習者 J：構成や日本語

【ホリ 4 点（4)、目・4 点（4)、構・結 4 点（4)、日 4 点（4)】

学習者 K：メタ言語を使ってみた

【ホリ 5 点（5)、目・内 4 点（4)、構・結 4 点（5)、日 5 点（5)】

学習者 L：メタ言語を増やす

【ホリ 5 点（4)、目・内 3 点（4)、構・結 4 点（4)、日 4 点（4)】

自由記述をみていくと、「4.2　評価への納得度」でも考察したように、高い評価点を得た場合は「満点だったから修正してない」のように推敲への動機が生まれないが、「点数が低いところを中心に見ていった」や「目的、内容のところが 4 だったので、目的に沿って話が進められていたかを修正しようとした」のように、評価点が低かった場合には推敲への動機が高まっていることが確認できる。

次に、「どんなことに気を付けて『修正』をしたか」への自由記述回答のテキストマイニング[(2)]の「ワードクラウド結果」を示す。「ワードクラウド結果」の見方は、スコアが高い単語が複数選び出され、その値に応じた大きさで図示されている。 単語の色は品詞の種類で異なっており、青色が名詞、赤色が動詞、緑色が形容詞、灰色が感動詞を表している。

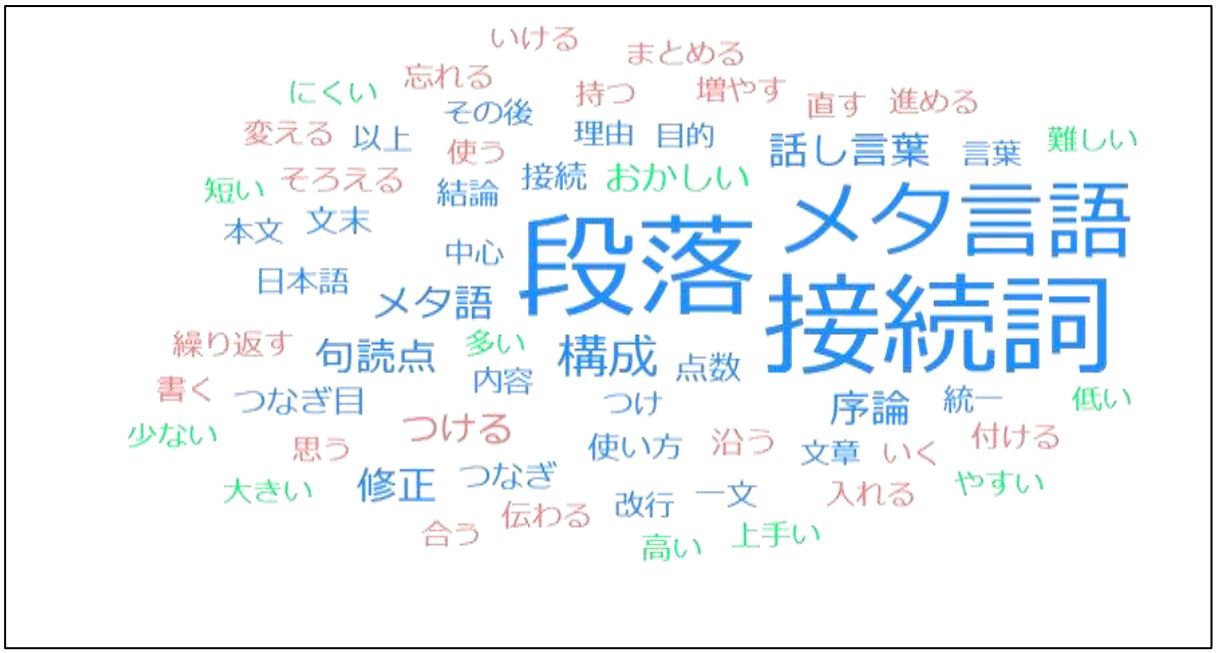

図 1　『修正』に対する自由記述のワードクラウド結果

「ワードクラウド結果」から、推敲の観点として「段落」「メタ言語」「接続詞」が大変強く意識されていることがわかった。それ以外では「構成」「話し言葉」「句読点」「序論」等にも意識が向いているようである。

GoodWriting Rater には、評価点を表出する機能の他に「テキスト情報」「メタ言語ハイライト」という機能がある。「テキスト情報」では、総文字数や総段落数、漢字率、第1段落の文字数/総文字数等、様々な情報が数値として表示される。「メタ言語ハイライト」では、入力した意見文のメタ言語部分が機能別に色分けされ、ハイライトされた状態で表示される。この2つの機能を比べると「テキスト情報」では数値が示されるだけだが、「メタ言語ハイライト」は視覚に訴える形で、文章中のメタ言語の使われ方がわかりやすく示されるという違いがある。

「4.2 評価への納得度」で示した通り、マルチ評価「目的・内容」「構成・結束性」で低い評価点が表出された場合、推敲への動機が高まる傾向がある。ここから考えられることは、その推敲への動機と「メタ言語ハイライト」による視覚的な刺激が相まって、推敲の観点として「段落」「メタ言語」「接続詞」が強く意識されるのではないかということである。

5. まとめと今後の課題

(1) GoodWriting Rater は日本人学部大学生（日本語母語話者）の書いた作文に対して、どのような評価を表出するのか。

マルチ評価「日本語」やホリスティック評価においては日本語母語話者であることが有利に働き高い点数が出やすいが、マルチ評価「目的・内容」や「構成・結束性」においては、ライティング能力の差が評価点の違いとなって表れることがわかった。

(2) 日本人学部大学生は、その評価にどの程度納得するのか。

評価への納得度に関しては、高い評価点が示された場合は納得し、一方、評価点が低かった場合は納得しづらいのか、評価の方法や基準、その評価点が導き出された理由などを知ろうとする動機が生まれることがわかった。

(3) GoodWriting Rater の日本人学部大学生への活用可能性はあるのか。

今回の GoodWriting Rater の試用に関して、意見文への評価点が高い場合は推敲への動機が生まれず、低い場合は推敲への動機が高まるものの、実際の推敲では「メタ言語ハイライト」のようなわかりやすい刺激に反応しがちであることがわかった。つまり、GoodWriting Rater の活用可能性という点から考えれば、たとえば、評価点が高い場合にもなぜ高く評価されたのか「テキスト情報」や「メタ言語ハイライト」といった機能を用いて分析する姿勢がほしいところであるが、「テキスト情報」や「メタ言語ハイライト」の活用については、日本人母語話者であっても学生個人の力では御しがたいと推測されることがわかった。GoodWriting Rater の活用可能性を高めるには、教師による学習支援や指導が必要であることがいえるだろう。

今回の調査報告では、開発過程にある日本語作文（意見文）の自動評価システムの試用実態の報告にとどまり、十分な分析はできなかった。今後は、数値を統計的に分析することや推敲前後の作文の変化について質的調査を実施する等、詳細かつ具体的な分析を行う必要があるだろう。さらに、GoodWriting Rater の持つ3つの機能を学習者が自律的に使

えるようになるためには、教師によるどのような支援や指導が有効かについても考えていきたい。

（影山陽子かげやまようこ・日本女子体育大学・kageyama@jwcpe.ac.jp）

謝辞

本研究は、科学研究費基盤研究（B）26284074「日本語ライティング評価の支援ツール開発：「人間」と「機械」による評価の統合的活用」（田中真理代表）の研究成果である GoodWriting Rater を利用して行われたものです。

注

1. 一部米国のデータも含まれている。
2. ユーザーローカル テキストマイニングツール（https://textmining.userlocal.jp/）を用いて分析を行った。

参考文献

石井雄隆・近藤悠介（2013）「英語学習者を対象とした自動採点システム―課題と展望」『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』, 1-11.

齋藤雪絵（2017）「自動採点システムを使った英語ライティング学習」『立教大学ランゲージセンター紀要』38, 63-74.

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』, 75-92.

田中真理・阿部新（2014）『Good writing へのパスポート―読み手と構成を考えた日本語ライティング』, くろしお出版

独立行政法人大学入試センター
＜https://www.dnc.ac.jp/daigakunyugakukibousyagakuryokuhyoka_test/＞（2019 年 2 月 15 日閲覧）

藤田彬・藤田央・田村良直（2012）「国語教育的評価項目を考慮した機械学習による日本語文章の自動評価と評価モデルの構築」『Journal of natural language processing』19(4), 281-301.

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』, 289-291.

ETS TOEFL＜https://www.ets.org/jp/toefl/ibt/scores/＞（2019 年 2 月 15 日閲覧）

GoodWriting.jp―読み手と構成を意識した日本語ライティング
＜https://goodwriting.jp/wp/＞（2019 年 2 月 15 日閲覧）