英語学習者が産出した英語ライティングの評価に関する考察 : 評価者はどのように第二言語ライティングを評価しているのか

成田, 眞澄

(Citation)
Learner Corpus Studies in Asia and the World,6:121-132

(Issue Date)
2024-03-20

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/0100487718

(URL)
https://hdl.handle.net/20.500.14094/0100487718

英語学習者が産出した英語ライティングの評価に関する考察

―評価者はどのように第二言語ライティングを評価しているのか―

成田　眞澄（津田塾大学）

A Consideration on the Evaluation of English Writing Produced by Learners of English

―How Raters Evaluate L2 Writing Performance―

NARITA, Masumi (Tsuda University)

概要

本稿は，第二言語ライティングの評価者の特性や評価ルーブリックが評価のプロセスにどのような影響を及ぼすのかを探る研究に向けた予備的な考察である。評価者の評価経験や分析的評価ルーブリックが評価のプロセスに及ぼす影響に関する主な先行研究を概観し，評価者による第二言語ライティングへのコメントを新たな分析対象として取り上げる。

キーワード

第二言語ライティング，ライティング評価，評価ルーブリック，多相ラッシュ分析

1. はじめに

規模の大きい英語能力試験においても，日々の教室における英語指導においても，英語学習者による英語でのスピーキングやライティングの能力をどのように評価するかという課題は，評価結果がもたらす影響とも相俟ってきわめて重要である。近年では，いずれの産出能力の評価においても評価の観点や尺度を記述したルーブリックやチェックリストの使用が推奨されるようになってきた（小泉・印南・深澤，2017；山下，2023）。同時に，これらの評価基準は，普段の授業のなかで英語教員が学習者の産出能力の伸びを測定してフィードバックを提供するために，あるいは学習者自身による自己評価やピアによるフィードバックを促すために使用することもできる（山下，2023）。

筆者が担当している英語ライティング科目においても，パラグラフライティングやエッセイライティング，サマリーライティング，リアクションペーパーといった英語ライティングの作成対象に応じたルーブリックを事前に用意し，使用している。学習者にもシラバスで共有している。英語ライティング科目を担当する複数の教員全員がプロセスアプローチによる指導のなかで適宜これらを使用して学生の英語ライティング能力を測定するとともに個別にフィードバックを提供している。しかし，こうした

英語ライティング指導を継続して実践するために，評価とフィードバックの提供に多大な時間を要するという負荷を評価者である担当教員に課してしまっていることは否めない。筆者は，ルーブリックに含まれる観点と尺度を理解した上で，学習者が産出した英語ライティングを読み，時に両者を見比べながら評点を決めるというプロセスを辿ることが多いが，そのあとに学習者へのフィードバックを準備する時間も必要となる。

2022 年度に筆者が所属する学部で実施した英語ライティング担当教員と学習者へのアンケート調査とインタビュー調査の結果によると，教員はライティング評価とフィードバック提供に要する負荷を軽減するための方策を探りつつもやはり多大な時間を要してしまうという負担に言及しているが，学習者は教員から得られるフィードバックをきわめて肯定的に捉えていることが明らかになった（Narita et al., 2023）。このような実態に直面し，ルーブリックを用いた英語ライティング評価のプロセスに関わる要因を探る必要性を考えるようになった（Greene et al., 1989; Barkaoui, 2010）。

本稿は，英語学習者が産出した英語ライティングの評価において，評価者の特性や評価に使用するルーブリックが評価のプロセスにどのような影響を及ぼすのかという研究に向けた予備的な考察である。

2. 第二言語ライティング能力の評価法と評価者

ライティング能力は，特定のライティングタスクや分析対象とする言語的特徴に応じて評価されることも多いが，一般的な能力試験や教室でのライティング指導において広く使用されているのは，全体的評価（holistic scoring）と分析的評価（analytic scoring）に大別できるであろう（Weigle, 2002）。この場合でも，目的やジャンルによって評価の観点や尺度，すなわちルーブリックの設計は変わりうる。たとえば，分析的評価のための種々のルーブリックについては，保田（2024）に詳述されている。

全体的評価と分析的評価のいずれにもメリットとデメリットがあり，評価の規模や目的によって適切に選ぶ必要がある。選択に際しては，Bachman & Palmer (1996) による「テストの有用性」という概念に基づく両者の比較（Weigle, 2002: 121）が参考になる。評価法と使用するルーブリックが決まると，ライティングサンプルを用いて（通常は複数人の）評価者に対する評価トレーニングが実施され，評価者間の信頼性を高めたのちに実際の採点が開始される。この評価プロセスが始まると，評点を決めるための相互作用が採点対象のライティングと評価ルーブリック，そして評価者の間に生じると考えられる。評価者は，ライティングとルーブリックの解釈に加え，評点を決める判断を必要とするため，主観性を排除することは難しい。評価のための訓練を行っていても，この主観性は評価者による評価行為のばらつき（variability）につながり，評点と評点に基づく合否や習熟度判定といった意思決定に影響を及ぼしうる（Eckes, 2011; Lumley, 2005）。

Barkaoui (2010) は，評価法の違いが評価者や評点の信頼性に及ぼす影響についての研究は従来から取り組まれているものの，評価者が評価ルーブリックに基づいてどのように評価を行なっているのかというプロセスそのものを探ることが必要であると述べている。たとえば，全体的評価法では，評価者はルーブリックに基づいて評価対象のライティングが有する言語的特徴を頭のな

かで検討しながら最終的にひとつの評点に決定するため，評価者にとって認知的な負荷は高くなることが見込まれる，としている（p.55）。確かに，分析的評価法ではルーブリックにある観点別に採点したのちに合計点を算出するという流れを辿ればよいだけかもしれない。しかし，後述するように評価者による評価行為はもっと複雑であり，認知的な負荷と大きく関係してくるように思われる。

次節以降では，評価者と評価ルーブリックの相互作用に関する先行研究をいくつか紹介し，評価者の特性の違いによる影響も含め，評価者の評価行為がいかにばらつくかということを明らかにする。さらに，評価に関わる複数の変数を同時に扱うことができる多相ラッシュ分析（Many-Facet Rasch Analysis）の有用性についても触れる。

3. 評価者と評価ルーブリックの相互作用

3.1 評価経験と評価法の違いによる影響

評価者の特性（母語や評価経験，誤りに対する許容度の違いなど）といった評価者要因が評価に及ぼす影響を調査した研究はこれまでも数多くなされてきた（Barkaoui, 2010; Cumming, 1990; Huang, 2009; Lumley, 2005）。たとえば，Cumming（1990）では，評価に関する専門知識を有し，評価経験が豊かな評価者と未熟な評価者との間には評価のプロセスに質的な違いがあることを指摘している。

評価者の特性と評価ルーブリックの双方を分析対象とした研究が Barkaoui（2010）である。評価者の特性として着目したのは評価経験の有無である。具体的には，思考発話法を用いて，第二言語としての英語（English as a Second Language: ESL）ライティングの評価を経験したことのない評価者 11 名と経験豊富な評価者 14 名との間には全体的評価ルーブリックあるいは分析的評価ルーブリックに基づく評価のプロセスにどのような違いがあるのかを調査した。分析結果（p.64）を以下の表 1 にまとめる。そして，類似点もあるとした上で，分析的評価ルーブリックを用いたほうが言語的適切性に対してより多くの注意が向けられたと報告している。

表 1

Barkaoui（2010）における評価者の評価プロセスの違い

評価者	全体的評価ルーブリックの場合	分析的評価ルーブリックの場合
評価経験有（14 名）	・書き手の状況を想定しながら，論拠や修辞法を評価する	・ルーブリックよりも頻繁に文章を読み返し，文章の長さや言語的特徴に注意を向ける
評価経験無（11 名）	・語彙や文構造などの言語的特徴をきめ細かく考慮してスコアを決める	・ルーブリックに従いながら評価を行うが，文章構成にも注意を向けるようになる

さらに，統計的な分析が加わると，評価者の特性よりも評価ルーブリックのほうが評価者の意思決定や注意を向けるライティングの要素に対する影響が大きいと結論づけている。評価の経験のない，あるいは経験が不足している評価者には分析的評価ルーブリックを用いた評価法が適しているのではないかと考察するとともに，評価ルーブリックに記述されている評価の観点・尺度の順番が評

価プロセスに影響を及ぼしうると示唆している。

英語を外国語として（English as a Foreign Language: EFL）学んでいる日本人大学生が産出した英語ライティングを対象として，多相ラッシュ分析を用いて評価者に生じるバイアス（偏り）を調査した研究に Schaefer（2008）がある。多相ラッシュ分析は，前述したように評価に影響を与えうる複数の変数（あるいは相）を同時に分析できるという利点があり，たとえば評価者と評価の観点の間に偏った評価傾向が見られるかどうかを調べることもできる（Barkaoui, 2013; Eckes, 2011）。Schaefer (2008)では，日本人大学生 40 名による英文エッセイを 40 名の英語母語話者（東京近郊で働いている英語指導助手）が分析的評価ルーブリックを用いて評価した。この評価ルーブリックは既存のルーブリックを参考にして独自に開発したものである。評価の観点として Content（内容），Organization（構成），Style and Quality of Expressions（表現のスタイルと質），Language Use（言語使用），Mechanics（機械的技術），Fluency（流暢性）の6つがあり，各観点には 1，2 文程度の簡潔な記述語（descriptors）と 6 つのレベル（Excellent, Good, Fair, Barely adequate, Poor, Very poor）が用意された。

こうして得られた評点に対して多相ラッシュ分析（特にバイアス分析）を行なった結果，評価者と評価の観点との間には，(1)「内容」および「構成」の評価が厳しい評価者は「言語使用」および「機械的技術」の評価は甘い傾向があった，(2)反対に，「内容」および「構成」の評価が甘い評価者は，「言語使用」および「機械的技術」の評価は厳しい傾向が見られた。さらに，評価者と書き手（実験参加者）との間には，優れた書き手に対してはそうではない書き手に対してよりも評価の厳しさにおいて偏りが大きくなる傾向が観察された。

上記のいずれの研究においても，評価者と評価ルーブリック，あるいは評価対象である第二言語ライティングとの間に評価の偏りが生じうることが明らかになった。このことは，第二言語ライティング能力の評価の過程において評価者の意思決定に影響を及ぼす要因を明らかにし，評価者に対する事前訓練に活かす必要があることを示している。

3.2 ESL Composition Profile が評価者に及ぼす影響

本節では，第二言語ライティング能力の分析的評価ルーブリックとして広く使用されている ESL Composition Profile (Jacobs et al., 1981)に研究者が独自に手を加えた修正版を用いて評価者のふるまいを調査した研究を紹介する。筆者も日本人大学生が産出した英語論述文を対象とする研究において ESL Composition Profile を何度も使用してきたが，実は評価者間信頼性（interrater reliability）が高かった場合とそうではなかった場合があり，この分析的評価ルーブリックが評価者によってどのように理解され，どのような評価のプロセスをもたらしうるのかを探ることは重要であると考えている。

まず，第二言語ライティング能力の分析的評価ルーブリックである ESL Composition Profile の概要を図 1 にまとめる。複数の観点から第二言語ライティング能力を評価できるため，測定時点での各英語学習者の強みと改善点が明らかになり，教員からは適切なフィードバックを英語学習者に提供することができる。しかし，同一の記述語（descriptors）に対して付与する評点に幅があり，評

価者の間で評点の解釈と付与にばらつきが生じ，評価の妥当性を損なうことになりかねない，と Chang et al.（2023）は指摘している（p. 2）。

図 1

ESL Composition Profile（Jacobs et al., 1981）の概要

評価の観点と評点割合	内容（30），構成（20），語彙（20），言語使用（25），機械的技術（5）
レベル分け	4 段階（EXCELLENT TO VERY GOOD，GOOD TO AVERAGE，FAIR TO POOR，および VERY POOR）
評価基準 （例.「内容」の評価）	EXCELLENT TO VERY GOOD（30-27）：knowledgeable ● substantive ● thorough development of thesis ● relevant to assigned topic GOOD TO AVERAGE（26-22）：some knowledge of subject ● adequate range ● limited development of thesis ● mostly relevant to topic, but lacks detail FAIR TO POOR（21-17）：limited knowledge of subject ● little substance ● inadequate development of topic VERY POOR（16-13）：does not show knowledge of subject ● non-substantive ● not pertinent ● OR not enough to evaluate

注： Weigle（2002）の Figure 6.3（p.116）に基づいて作成

評価者が ESL Composition Profile（研究者が手を加えた修正版）を用いて ESL ライティングを評価する際の評価者の認知プロセスを分析した研究のひとつに Winke & Lim（2015）が挙げられる。この研究では，過去に大学で実施されたプレイスメントテストの結果によるグループ分け（高得点グループ，中得点グループ，および低得点グループ）とエッセイの長さ，プロンプトの違いを考慮に入れて選択された 40 のエッセイに対して 9 名の評価者が事前訓練を受けたのちに評価をするが，その採点中の視線の動きや視線が停留する箇所や時間，頻度を計測している。パソコン画面上に提示されたルーブリックのどの部分をどのくらいの時間注視しているかを測定できる装置を使用している。評価対象のエッセイと採点記録表はパソコンの手前の机上に用意された。視線計測データの分析結果を以下にまとめる。

(1)評価者は，ルーブリックに記載されている観点の順に評価を行う傾向があった。すなわち，使用されたルーブリックに記載されている観点の順番は，「内容」，「構成」，「語彙」，「言語使用」，「機械的技術」であったが，評価もこの順番で行われた。

(2)ルーブリックの最後の観点である「機械的技術」の評価にかける時間が少なかった。

(3)より多くの注意を向けた観点に対する採点の評価者間信頼性は高くなる傾向があった。最も注意を向けた「構成」の採点における評価者間信頼性が最も高い数値を示し，最も注意を向けなかった「機械的技術」の採点における評価者間信頼性が最も低い数値であった。

(4)評価者間信頼性の高い評価者は似たような評価プロセス（ルーブリックにおける注視のパターンや点数調整の柔軟性）を経るが，評価者間信頼性の低い評価者の注視パターンには個人

差が顕著に見られた。

分析的評価ルーブリックにおいて評価の観点が記載される順番が評価者の評価行為や評価者間信頼性に影響を及ぼしうるという結果は，ルーブリックの設計に関する研究を推し進める必要があることを示している。さらに，著者たちは，視線計測データは評価者がなぜ特定の観点により多くの注意を向けたのかを説明してくれるものではないという限界を認めつつも，たとえば刺激想起インタビューと組み合わせることで評価者がとる行動の理由を探ることができるのではないかと考察している。

次に，ハワイ大学が実施する英語集中講座のプレイスメントテストで収集された英文エッセイを対象として，独自に手を加えた ESL Composition Profile を用いて採点し，多相ラッシュ分析を行った最近の研究に Chang et al.（2023）がある。この研究では，(1)ESL Composition Profile から「機械的技術」の観点を排した 4 つの観点に均等に 6–25 点（レベルの数はオリジナルと同様に 4 つ）を割り当てて評価した結果，(2)(1)の観点ごとに具体的な点数ではなく 4 段階評価とした結果，さらに(3)(2)の評価法を直近のエッセイ評価で試行した結果が報告されている。(1)と(2)における評価法の違いを図 2 に示す。

図 2

観点ごとに評価法(1)あるいは(2)を用いて評価する場合のルーブリックの例（「内容」の評価）

	(1) レベルと配点	(2) レベルと段階	評価基準
観点「内容」	21–25	4	Essay clearly addresses topic ● Ideas are developed thoroughly ● Essay reflects substantive thought ● No extraneous material
	16–20	3	Essay mostly focused on topic ● Expresses a few advanced ideas ● Some details and reasons included, though thesis not fully developed
	11–15	2	Essay minimally addresses the topic (at the surface level) ● Development of ideas is not complete ● Lacks detail and support
	6–10	1	Essay does not adequately address the topic ● Ideas are either non-substantive or not pertinent ● OR Not enough to evaluate

注：この表は Chang et al.（2023）の Figure 1（p.4）および本文の記述に基づいて作成

上記(2)の評価法は，(1)の評価において評価者による評点が完全に一致した割合がきわめて低い数値（14%）であったことからシミュレーションを経て 4 段階評価に改訂したものだが，評点の完全一致率は 50%に向上した。多相ラッシュモデルが予測した一致率は，それぞれ 15%と 56%であった。数値としてはまだ低いと言えるが，4 段階評価とすることにより受験者の英語ライティング能力の違いがテスト実施者の意図に即して測定できるように改善されたこと（ラッシュモデルへの適合度

や各得点の確率曲線に基づく）が多相ラッシュ分析により明らかになった。その後の上記(3)で実施した試行においても良好な結果が得られた。

このような結果から，Chang et al.（2023）はオリジナルの ESL Composition Profile のように評点の幅を細かく設定することは教室環境での英語学習者へのフィードバックという観点からは有益であると考えるが，プレイスメントテストのような場合には意味のある解釈につながるのかどうか―得られた評点がライティングの質の違いを適切に捉えられるのかどうか―疑問視している。

3.3 第二言語ライティングへの評価者コメント

教室環境では，学習者が産出した第二言語ライティングに対して指導教員が評価コメントを提供することは珍しくないだろう。一方，能力試験や実験環境では，評価者が評価ルーブリックを用いて第二言語ライティングを評価するわけであるが，実験実施者からの特別な指示がないかぎりは評点の付与のみで終わることが多いだろう。

しかし，2023 年 10 月に International Corpus Network of Asian Learners of English（ICNALE）プロジェクトから公開された Global Rating Archives（GRA）（Version 2.0）には評価者によるコメントが長所（strong points）と短所（weak points）に分けて記載されており，評価者が個々の英文エッセイのどんな言語的特徴に注意を向けたのか，といった新しいアプローチでの評価者行為の分析が可能になった。

Ishikawa（2023）によると，GRA では ICNALE プロジェクトで収集した産出データから英語での 140 のスピーチと 140 のエッセイを抽出し，L1 や出身地，職業において異なる背景を持つ合計 80 人の評価者が全体的評価と分析的評価の両方を行い，評点に加えてコメントも付与している。分析的評価は，ESL Composition Profile での評価経験に基づき，第二言語による産出の質に影響を与えると考えられる 3 つの基本的な側面から合計 10 の観点に基づき，それぞれ 10 点満点で採点を行うという独自の方法を採用している（表 2）。特に，これまでの評価ルーブリックにはなかった attitude（「態度」）という側面を取り入れたことと評価者の評価プロセスを統制する仕掛けを組み入れたことは，ICNALE GRA を斬新で特別なコーパスにしている。

表 2

ICNALE GRA における分析的評価の構成

Language	*Content*	*Attitude*
Intelligibility	Comprehensibility	Willingness to communicate
Complexity	Logicality	Involvement
Accuracy	Sophistication	
Fluency	Purposefulness	

注：Ishikawa（2023）の TABLE 3.12（p.65）に基づいて作成

評価者のバックグラウンドの多様性に着目すると，同一の評価ルーブリックを使用していて

も職業の違いが評価プロセスに影響をもたらすのかどうかを調査することもできる。たとえば，英語教員と英語以外の科目を教えている教員の比較や英語教員とビジネスパーソンの比較を行うことにより，評価プロセスに影響を及ぼす評価者要因が，しかもこれまでの研究で気がついていない要因が見えてくるかもしれない。

筆者は，予備的な調査として，ICNALE GRA に含まれる EFL 環境の学習者が産出した 120 のエッセイに着目し，評価者 80 名から無作為に抽出した英語教員 17 名とビジネスパーソン 17 名によって付与された長所（strong points）と短所（weak points）のコメントを比べてみた。AntConc（Anthony, 2022）による頻出単語と KWIC の表示により，これらのバックグラウンドの異なる評価者によるコメントの大まかな傾向を調べた。コメントの記述には類似した表現が見つかるが，先に記述したコメントを流用できる場合には再利用されたものと考えられる。

長所のコメントとして，ビジネスパーソンは「主張を裏付ける記述がなされていること」（図 3）を，英語教員は「考えが論理的に述べられていること」（図 4），「エッセイが明瞭で理解しやすいこと」に言及したものが多く見られた。短所のコメントでは，いずれのタイプの評価者も文法に関するコメントが目立つものの（図 5 および図 6），英語教員の場合には文の構造（図 6）や理解のしやすさ，文と文のつながり（図 7）に難点があることへの言及が多く見られた。

今後，評価者によるコメント記述に対する詳細な分析が必要であるが，評価ルーブリックに記載された評価基準に影響を受けたコメントになることが予想されるため，Cai & Yan（2023）のように評価ルーブリックとコメント記述の言語的な関係性についても探ることができるようになる。

図 3

AntConc（Version 4.2.0）による KWIC 表示（ビジネスパーソンによる長所コメントの一部）

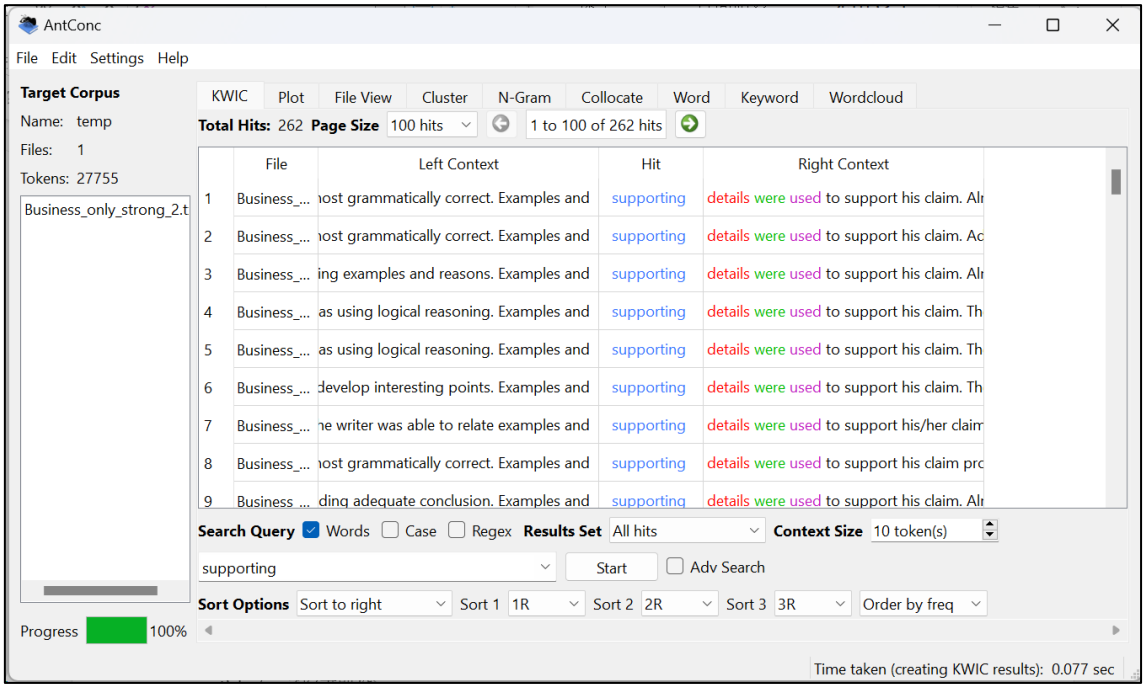


図 4

AntConc（Version 4.2.0）による KWIC 表示（英語教員による長所コメントの一部）

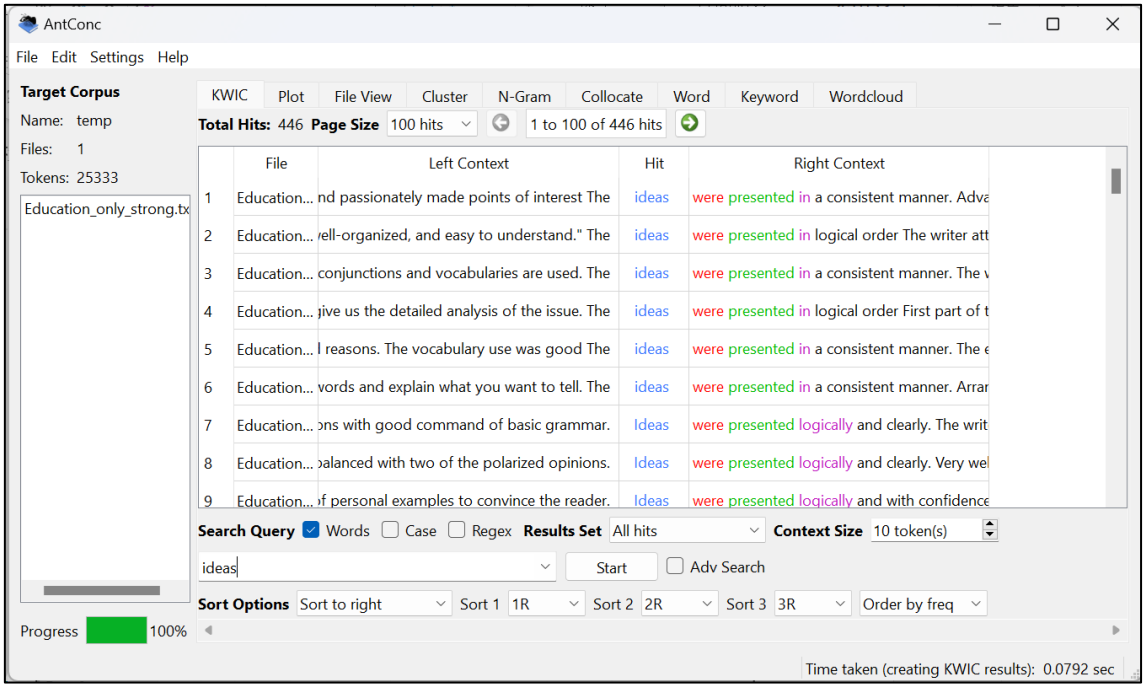


図 5

AntConc（Version 4.2.0）による KWIC 表示（ビジネスパーソンによる短所コメントの一部）

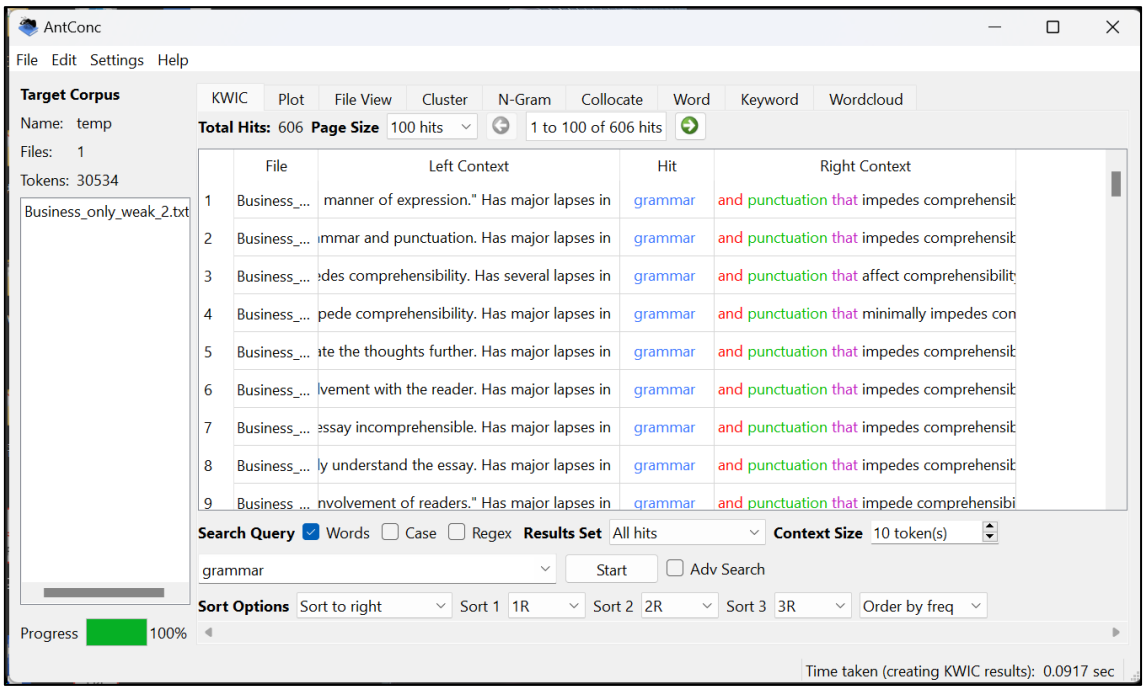


図 6

AntConc（Version 4.2.0）による KWIC 表示（英語教員による短所コメントの一部）

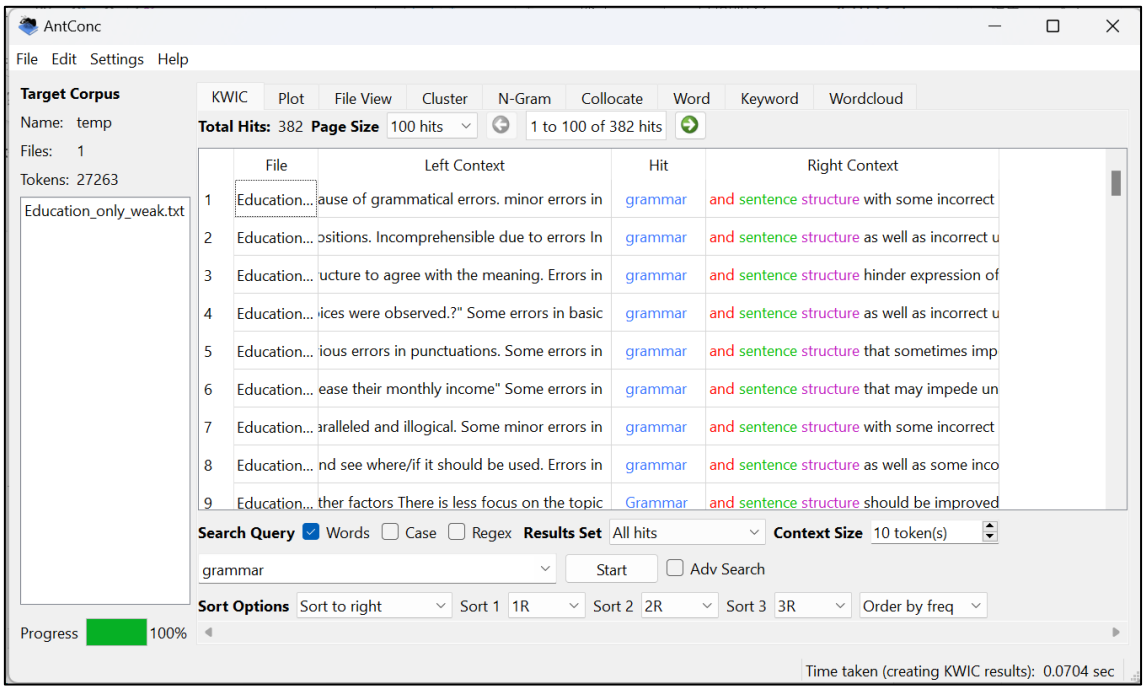
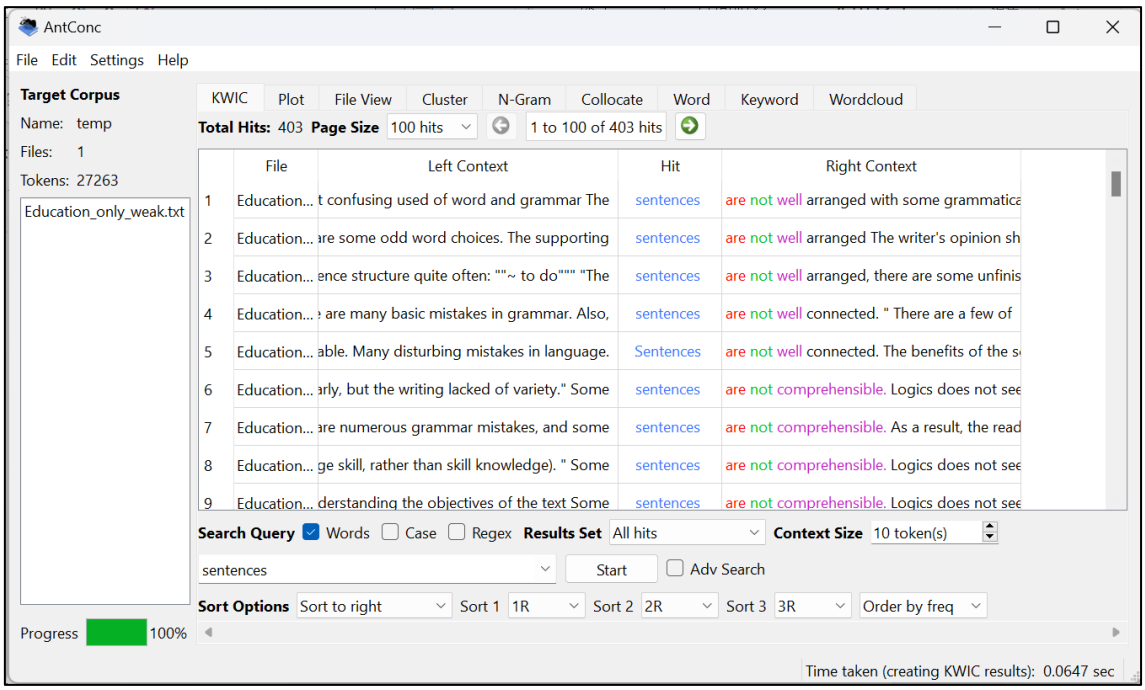


図 7

AntConc（Version 4.2.0）による KWIC 表示（英語教員による短所コメントの一部）



4. まとめ

現在，筆者は ICNALE プロジェクトから公開されている日本人英語学習者による英文エッセイデータ（の一部）に対して，英語を母語とする英語教員 4 人が ESL Composition Profile（オリジナル版）を用いて評価した結果を分析している。多相ラッシュ分析によって算出された評価者間信頼性の数値は 17.2%と低く，Chang et al.（2023）のように 4 段階評価に変えてみると 45.6%に上がったが依然として低い。さらに，バイアス分析により，評価者と評価の観点との間に見られる偏り―ルーブリックにあるどの観点がどの評価者によってより厳しく，あるいはより甘く評価されているのか―が明らかになった。これにより，第二言語ライティングの指導と評価において，評価の結果としての評点だけではなく，ライティング（つまり書き手）と評価者，評価ルーブリックがお互いにどのように影響しあっているのかという評価のプロセスを探る重要性を痛感している。

本稿で紹介した先行研究からも，評価者の特性や評価ルーブリックが評価のプロセスに影響を与えうることがわかる。人間による評価だからこそ受ける影響であるが，評価のプロセスをさまざまな手法により探ることの意義は大きいと考える。その結果は，評価者への評価訓練に加え，学習者へのフィードバックにも活かせるだろう。

引用文献

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Waseda University. Available from https://www.laurenceanthony.net/software

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74.

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. In A. Kunnan (Ed.), *The companion to language assessment* (Vol. III: Evaluation, methodology, and interdisciplinary themes, Part 10: Quantitative analysis) (pp.1301–1322). John Wiley & Sons.

Cai, H. & Yan, X. (2023). Triangulating NLP-based analysis of rater comments and MFRM: An innovative approach to investigating raters' application of rating scales in writing assessment. *Language Testing*. https://doi.org/10.1177/02655322231210231

Chang, Y.-T., Choe, A. T., Holden, D., & Isbell, D. R. (2023). Making each point count: Revising a local adaptation of the Jacobs et al. (1981) ESL COMPOSITION PROFILE rubric. *Language Testing*. https://doi.org/10.1177/02655322231217979

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and*

evaluating rater-mediated assessments. Peter Lang.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies*, *5*(1), 1–17.

Ishikawa, S. (2023). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Jacobs, H., Zinkraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Newbury House.

小泉利恵・印南洋・深澤真(2017). (編著)『実例でわかる英語テスト作成ガイド』大修館書店.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Peter Lang.

Narita, M., Okuwaki, N., & Gray, G. (2023). Developing critical thinking skills: A pedagogical inquiry into Japanese learners of English. In T. Muller, J. Adamson, S. Herder, and P. S. Brown (Eds.), *Re-Envisioning EFL Education in Asia* (pp.1–17). International Teacher Development Institute.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 463–492.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 38–54.

山下美朋(編著)・河野円・長倉若・峰松愛子・山岡憲史・山中司(著) (2023).『英語ライティングの指導:基礎からエッセイライティングへのステップ』三修社.

保田幸子(2024).『「書く力」の発達:第二言語習得論と第二言語ライティング論の融合に向けて』くろしお出版.