資　料

日本大学生産工学部研究報告B
2019 年 6 月 第 52 巻

英語ライティング指導のための自動フィードバックシステムの開発に向けて

小林雄一郎*，石井雄隆**

Toward an Automated Feedback System for English Writing Instruction

*Yuichiro KOBAYASHI**，*Yutaka ISHII***

In recent years, the importance of feedback for individual learners has drawn attention in language teaching, and has accelerated the research on diagnostic feedback. Furthermore, there is an increasing tendency for language teachers to use automated essay scoring system such as e-rater. In response to such a trend, the purpose of the present paper is to explore the design of automated feedback system for English language teaching. By reviewing the theories of language testing and writing instruction, this paper discusses the specifications of automated essay evaluation system in EFL (English as Foreign Language) settings. This paper also proposes a set of linguistic features that can be computationally analyzed using natural language processing techniques for the development of computer-based feedback system.

Keywords: Automated Scoring, Automated Feedback, Computer Assisted Language Learning

1．はじめに

近年，個々の学習者に特化したフィードバックの重要性は，言語テストの分野で注目されており，最適な診断的フィードバックの方法が理論的・実践的に模索されている。また，言語習得の分野では，学習者の語彙や文法の誤りに対する訂正フィードバックの効果が盛んに研究されている。しかし，従来のフィードバック研究では，研究の方法論やデザインなどの不備によって，同じ言語形式を対象とした研究であっても，結果が一致しないことが多い（大関，2015）[1]。そして，冠詞や不規則動詞などの非常に限られた数の言語項目のみがフィードバック研究の対象とされてきた。それに加えて，フィードバックの方法に関しても，評価者によって大きく異なることが明らかにされている。さらに，言語教育の現場に目を向ければ，クラスサイズの問題もあって，教員が学習者全員に丁寧なフィードバックを繰り返し提供するのは困難である。

現在，評価者によって評価が異なるという問題を解決し，教育現場の負担を軽減するという目的では，自動採点（automated scoring）の技術が大きな注目を集めている（Shermis & Berstein, 2003；Shermis & Berstein, 2013）[2), 3)]。すでに米国では，TOEFL iBT（Test of English as a Foreign Language Internet-Based Test）などの英語検定試験や，GMAT（Graduate Management Admission Test）や MCAT（Medical College Admission Test）などの大学院進学試験にライティングの自動採点が導入されている。しかし，e-rater や IntelliMetric などの既存の自動採点システムでは，全体的なスコアと，少数の言語項目に関するスコアやフィードバックが返されるだけで，言語習得研究や教育現場に

*日本大学生産工学部教養・基礎科学系助教
**早稲田大学大学総合研究センター助手

とって十分な情報が提供されているとは言えない。従って，スコアだけでなく，それ以外の様々な情報も提示する自動フィードバック（automated feedback）システムの実現は急務である。

本論の目的は，これまでの自動採点研究およびフィードバック研究に基づき，教育現場で実際に活用できる自動フィードバックの方法について検討することである。

2．自動採点システムの歴史と現状

近年，自動採点が大きな注目を集めている背景には，⑴教育環境におけるコンピュータの整備，⑵データ解析技術の発達，⑶グローバル化による英語学習者の増加，などの要因がある。まず，コンピュータを用いたテストの利点として，従来のペーパーテストと比べて，テストの配布や回収の自動化ができること，大量の答案の管理が容易になること，事前に用意したアイテムバンクから問題をランダムに出題できること，採点が自動化もしくは半自動化されること，学習者の言語能力に合わせた適応型のテストが実施できること，遠隔地での受験が可能になること，などが挙げられる（Ueno, 2005）[4)]。そして，自然言語処理や機械学習などの人工知能技術の発達によって，多肢選択問題や空所補充問題だけでなく，構成的応答（自由回答）の自動採点の研究も可能になった。

現在，ライティングの自動採点システムは，いまだ発展段階にあり，人間の評価者の完全な代替となってはいない。しかしながら，グローバル時代に対応した英語学習者の養成，大学進学率の上昇などの社会的要因によって，多くの学習者の言語能力を効率的かつ客観的に測定するための技術が強く求められている。実際，韓国では，KICE（Korea Institute for Curriculum and Evaluation）という国立機関で韓国人学習者の英語力を自動採点するための研究が進められており（Shin, Min, Park, Jung, Joo, & Kim, 2013）[5)]，日本においても，2020 年度から開始される大学入学共通テストで用いられる民間の資格・検定試験の中には自動採点を導入する予定のものが存在する[注1)]。

世界最初の本格的な英文自動採点システムは，Ellis Batten Page が 1960 年代半ばに開発した PEG（Project Essay Grade）である。このシステムでは，平均文長やパラグラフの数といった言語情報を手がかりに，重回帰分析という統計手法を用いて，学習者のライティング能力を推定している。PEG は，1990 年代に大幅に改訂され，ウェブ上での受験も可能になった（Page, 2003）[6)]。PEG のバージョンアップとほぼ同時期に，Vantage Learning 社が IntelliMetric という自動採点システムを公開した。このシステムは，意味的，統語的，談話的な言語情報を含む 300 以上の評価項目に基づき，ニューラルネットワークなどの人工知能技術を用いて，ライティングの評価を行っている（Elliot, 2003）[7)]。また，Thomas K. Landauer たちが開発した IEA（Intelligent Essay Assessor）は，LSA（latent semantic analysis）という手法を用いて，ライティングの形式面だけでなく，内容面も自動採点しようとしている（Landauer, Laham, & Foltz, 2003）[8)]。この手法は，使用語彙の頻度情報に基づいて，複数のライティングの内容がどの程度類似しているのかを数学的に測るものである。そして，現在最も有名な英文自動採点システムは，TOEFL や TOEIC（Test of English for International Communication）などのテストを運営している ETS（Educational Testing Service）が開発した e-rater である。このシステムでは，最先端の自然言語処理技術を駆使し，語彙，統語，談話的情報といったライティングの様々な側面を定量的に評価している（Burstein, 2003）[9)]。さらに，このシステムは Criterion というウェブベースのライティング支援ツールにも実装されており，教育現場で広く活用されている（Liao, 2016）[10)]。

言語テストを開発する場合には，「信頼性」と「妥当性」が重要となる。まず，テストの信頼性とは，同じ学習者に対して，同じ条件で同じようなテストを行った場合，同じ結果が得られる程度である。そして，テストの妥当性とは，そのテストで測定しようとしている能力（構成概念）を正しく測定できている程度である。従来，機械による自動採点は，人間による評価と比べて信頼性が高く，妥当性が低いとされてきた（Williamson, 2013）[11)]。それに加えて，人間による評価が多くの問題を抱えていることも，古くから指摘されてきた（Bejar, Williamson, & Mislevy, 2006）[12)]。いかに熟練した評価者であったとしても，英文における顕著な特徴に引きずられて他の特徴についての評価が歪められたり（ハロー効果），直前に読んだ英文が評価に影響を及ぼしたり（シークエンス効果），評価尺度の中心に評価が引きつけられたりすることもある（中心化傾向）。また，人間の評価は，長時間の作業による疲労の影響を受ける場合もあるため，信頼性が低くなりがちである（Ling, Mollaun, & Xi, 2014）[13)]。それに対して，機械による評価は，同一の英文に対しては常に同じ，一貫した結果を与える。その一方で，自動採点システムが人間のように正しく判定することは不可能である，という批判が繰り返し投げかけられてきた（Ericsson & Haswell, 2006）[14)]。実際，ライティングの一貫性や言語の創造的な使用など，現在のデータ処理技術による自動採点が難しい面は存在する（Higgins, Ramineni, & Zechner, 2015）[15)]。しかし，これは必ずしも技術的な問題ではない。言語を自動採点する場合，人間の評価者と同じ評価項目を用いることが理想ではあるが，専門的な訓練を受けた評価者であったとしても，自

分の評価基準に関する全てを言語化できる訳ではない（Attali, 2013）[16]。さらに，人間による評価も，彼らが批判する機械の評価と同様に，総語数や異語数といった言語の形式的な面や量的な情報の影響を色濃く受けているという報告もある（e.g., Kobayashi & Abe, 2016）[17]。また，自動採点の妥当性に関する研究は，妥当性そのものに関する研究の趨勢に大きく影響されてきた（Xi, 2012）[18]。従って，自動採点の妥当性についての結論を下すには，さらなる研究を積み重ねていく必要がある。

英文の自動採点システムを実装するには，⑴評価項目リスト，⑵学習者データ，⑶統計処理プログラム，の三つが必要となる。第一に，英文の自動採点を行う場合は，書き手のライティング能力を正確に測定するための言語的特徴のリストを策定しなければならない。しかし，前述のように，人間の評価者が用いている評価項目を全て把握することは，極めて難しい。従って，自動採点システムを実装するにあたっては，書き手のライティング能力と関連性があると思われる言語項目を可能な限り網羅的に検討する必要がある。因みに，既存の自動採点システムで広く使われている言語項目としては，統語に関わる指標（平均文長，T-unit の数，品詞 *n*-gram など），語彙に関わる指標（平均単語長，語彙多様性，語彙レベルなど），談話に関わる指標（談話標識の数，代名詞の数など）などがある。これらに加えて，文法的誤りを分析するシステムも存在するものの，現状の技術で自動検出可能な誤りは極めて限られている（文法的誤りの自動検出に関しては，本稿の 3.2.2 節を参照）。さらに，誤りを分析対象とする場合は，何を誤りとみなすか，回避などの方略をどう評価するか，なども考慮しなければならない。第二に，あらかじめ策定した評価項目リストに基づいて，異なるライティング能力を持つ学習者が書いた英文を比較する。多くの場合は，書き手の習熟度の情報が付与された学習者コーパスが用いられる（Higgins, Ramineni, & Zechner, 2015）[15]。また，コーパスに付与される習熟度の情報は，CEFR（Common European Framework of Reference for Languages）のレベル，TOEFL や TOEIC のような英語検定試験のスコアなどであることが多い。そして，評価項目とする統語的情報や談話的情報の分析にあたっては，品詞情報付与や構文解析といった自然言語処理の技術が活用される。第三に，個々の英文から集計した評価項目に関する値を用いて，書き手のライティング能力と個々の評価項目がどのような関係にあるかを統計的に記述する。そして，数学的に特定された「評価項目 *X* は，ライティング能力と正比例の関係にある」，「評価項目 *Y* は，初級者と中級者の弁別に有効である」といった無数のパターンから自動採点のためのプログラムが作成される。このようなパターンの抽出に用いられる統計手法には，重回帰分析，*k* 最近傍法，ベイズ判別法などがある（Larkey & Croft, 2003）[19]。

3．自動フィードバック・誤り訂正研究の動向

3.1　ライティングにおけるフィードバック研究の概観

本項では，自動フィードバックの必要性について論じる。自動フィードバックの研究動向を概観する前に，外国語教育研究や応用言語学の領域でどのようなフィードバック研究が行われてきたかを紹介する。Biber, Nekrasova, and Horn（2011）[20]では，これまでに刊行されたフィードバック研究のメタ分析を行っている。そして，2000 年から 2004 年の約 5 年の間に 100 件のフィードバック研究が行われていたと報告している。

フィードバック研究が盛んに行われた背景には，John Truscott と Dana Ferris の間で行われた論争が存在する。Truscott（1996, p. 327）[21]は，「第二言語ライティングのクラスにおいて，文法の誤り訂正をやめるべきだ」と主張した。その理由として，多くの研究が，⑴フィードバックを非効率的で役立たないと示しているということ，⑵理論的，実践的な理由において，フィードバックが非効率的であることが想定できるということ，⑶悪影響をもたらすことがあるということ，の三点を挙げている。

それに対して，Ferris（1999, p. 1）[22]は，Truscott（1996）の議論の共通点と相違点を議論しながら，Truscott が依拠している研究を調査した上で，Truscott の結論は「時期尚早で，あまりにも強すぎる」と反論をした。この論争は，様々なライティング・フィードバック研究が行われる契機となった。

フィードバック研究では様々な種類のフィードバックが扱われ，フィードバック方法の違いによる教育効果が検討されてきた。Ellis（2009）[23]は，Written Corrective Feedback（WCF）における教師のフィードバックを六種類に分類している。

1） Direct CF（corrective feedback）
2） Indirect CF
3） Metalinguistic CF
4） The focus of the feedback
5） Electronic feedback
6） Reformulation

Direct CF は，教師が学習者に正しい形を指摘するフィードバックを指す。Indirect CF は，エラーの箇所を示すものの，正しい形を示さないフィードバックである。Metalinguistic CF は，誤りの箇所にメタ言語的な説明をするタイプのフィードバックであり，エラーコー

ド（例えば，冠詞の誤りに *art*，前置詞の誤りに *pre* と記すなど）が代表的なものである。The focus of the feedback には，Unfocused CF と Focused CF の二種類が存在する。前者は学習者の誤りを全て修正することを指し，後者は修正する誤りのタイプを焦点化するタイプのフィードバックを指す。Electronic feedback は，教師が誤りを示し，正しい語法の例を示しているコンコーダンスラインへのハイパーリンクを示すフィードバックを指す。Reformulation は，学習者が産出したプロダクトに対して，元の内容を維持しながら，できるだけ学習言語の母語話者に近づけるように内容を再構成することを指す。

近年のフィードバック研究は，多様な展開を見せている。具体的には，実験デザインではなく，生態学的なデザインのフィードバック研究が行われてきている（e.g., Han, in press）[24]。たとえば，Amano（2018）[25]は，教員が原稿に書き込むフィードバックの種類を選択する権利を学生に与えた試みについて報告している。この研究によると，学習者が教師から受けたいフィードバックは多様であり，最初のドラフトとそれを修正したドラフトではそれぞれ異なるフィードバックを組み合わせるのが効果的であると学習者は感じている。具体的には，最初のドラフトでは，Direct CF をもらい，それを修正したドラフトでは，Question feedback，すなわち内容についてコメントをもらうフィードバックを好むと報告されている。学習者の選好を調査するフィードバック研究はこれまでも行われてきたが，学習者の好みに応じてフィードバックを行う研究はユニークな視点であり，自動フィードバックシステムの実装にとっても示唆的である。また，Han and Hyland（2018）[26]は，感情とフィードバックの関わりについても検討している。この研究によると，WCF がもたらすのは必ずしも否定的な感情ばかりでなく肯定的な感情でもあることや，WCF に対する感情的な反応は動的であって変化し得る。これらの点も，自動フィードバックについて検討する際には重要である。

本項の内容をまとめると，これまで多くの実験的なアプローチにより，様々なライティング・フィードバック研究の知見が蓄積されてきたが，近年では生態学的なデザインを採用したアプローチの研究も蓄積され始めており，それらの知見を基に自動フィードバックについて検討する必要がある。

3.2　自動フィードバックの現状

3.2.1　外国語教育・応用言語学の観点から

本項では，外国語教育・応用言語学の領域で行われてきた自動フィードバックについて論じる。最初に自動フィードバックの研究が出版されたのは 1980 年代であり，Grammar Writer's Workbench などがその先駆けとして知られている。このシステムでは，規則に基づいて文法的誤りを検出するアルゴリズムが用いられていた（Leacock, Chodorow, & Tetreault, 2015）[27]。1990 年代半ばになると，規則に基づくアプローチから統計に基づくアプローチへと少しずつ移行していった。この歴史的経緯については，自然言語処理の技術が発達していく過程と密接な関係があり，辻井（2012）[28]などに詳しい。

Dikli（2010）[29]は，自動採点システムによるフィードバックと教師によるフィードバックの長所と短所を**表 1** のようにまとめている。

表 1 にあるように，自動採点システムは，教師よりも短い時間で一貫したフィードバックができると言われている。しかしながら，自動採点システムが誤ったフィードバックを与える可能性があることも指摘されている（自動採点システムによるフィードバックの精度については，次項で言及する）。また，WCF と AWCF（Automated Written Corrective Feedback）には，下記の三つのような問題点が存在し得る（Ranalli, 2018）[30]。一つ目は，学習者にとって有益な情報量の違いが教育的な配慮というよりも技術的な限界によって決定されるかもしれないという点である。二つ目は，自動採点システムによる誤

表 1　自動採点システムと教師によるフィードバックの長所と短所（Dikli, 2010 に基づく）

	自動採点システム	教師
長所	・即座に採点とフィードバックができる ・一貫した採点と体系的なフィードバックができる ・人間が採点する手間がかからない ・多くのライティングを採点し，フィードバックを与えることができる	・具体的なフィードバックができる ・必要な情報だけを与えることができる ・人と人の関わり合いが存在する ・個人に合わせたフィードバックができる
短所	・人と人の関わり合いが存在しない ・一般的で冗長なフィードバックになりがちである ・過度なフィードバックを与えることがある ・誤ったフィードバックを与えることがある ・技術的な問題が生じる可能性がある	・人間が採点する手間がかかる ・採点とフィードバックが主観的になりがちである ・多くのライティングに対応する時間が必要となる ・一貫したフィードバックをしそこなう可能性がある

り検出の不正確さが学習者によるフィードバックの活用に悪い影響を与えるかもしれないという点である。三つ目は，個人差をほとんど（あるいはまったく）考慮しない自動フィードバックの汎用的な性質に限界がある点である。

3.2.2　自然言語処理の観点から

前項では，外国語教育研究の観点から AWCF について論じた。続く本項では，自然言語処理の観点から AWCF について論じる。自然言語処理の分野において，AWCF に最も近い研究領域は，文法誤り検出・訂正である。文法誤り検出・訂正には，以下のような三種類のタスクが存在する[注2)]。一つ目は，文法誤り検出であり，誤りを検出することを目的とするタスクである。二つ目は，文法誤り訂正であり，誤りを検出せずにそのまま正しい形に修正するタスクである。三つ目は，文法誤り検出・訂正であり，誤りを検出したうえで正しく訂正した形を示すタスクである。この中で自動フィードバックと最も関係が深いのは三つ目のタスクであるが，二つ目のタスクを indirect CF ととらえることも可能である。

これら三つのタスクでは，大きく分けて二つのアプローチが採用されている。一つ目は，規則に基づくアプローチである。たとえば，主語と動詞の一致に関する訂正などは，多くの場合，人手で設定した規則に基づく検出が可能である（永田，2014）[31)]。二つ目は，検出規則をコーパスから自動抽出するアプローチであり，*n*-gram などの言語モデルや機械学習が活用される。たとえば，言語モデルや機械学習に基づくアプローチは，前置詞誤りなどの訂正に向いており，大量のデータに基づいて正用と誤用を判定する際に有益である（Sakaguchi, Hayashibe, Kondo, Kanashiro, Mizumoto, Komachi, & Natsumoto, 2012）[32)]。

文法誤り訂正に対しては，様々な shared task が行われている。shared task とは，共通のデータセットに対して，それぞれが開発したシステムによる誤り検出を行い，その精度を競うコンペティションである。その主要なものとして，海外では Helping Our Own 1，Helping Our Own 2, 2013 conference on Computational Natural Language Learning（CoNLL 2013），2014 conference on Computational Natural Language Learning（CoNLL 2014）などが行われ，日本でも Error Detection and Correction Workshop 2012 などが開催された。

前述の主語と動詞の一致に関する訂正，前置詞誤り検出・訂正のほかには，冠詞誤り訂正（Rozovskaya, Chang, Sammons, & Roth, 2013）[33)]，時制誤り検出・訂正など（Tajiri, Komachi, & Matsumoto, 2012）[34)]の研究が行われている。

文法誤り検出・訂正の結果を評価するにあたっては，全体の誤りをどの程度検出できたかを表す再現率（recall）と，検出した誤りの中で実際にそれが誤りであった割合を示す適合率（precision）を計算し，再現率と適合率の調和平均である *F* 値を用いることが多い。しかしながら，これらの評価指標はコーパスサイズによって結果が異なるため，容易に一般化することができない。また，文法誤り検出・訂正については，人間がコーパスに文法誤りに関する情報を付与したデータを正解データとして用いるが，これらの人手による文法誤り情報の付与も正解が一義的に定まらないことも多い。

自動採点では，誤りの情報を使わずに，次節で紹介するような言語情報を使うだけでも一定の精度が得られることが知られている（e.g., Kobayashi & Abe, 2016）[17)]。しかしながら，教育現場，あるいは学習者自身からのニーズとして，ライティングにおける誤りの訂正が求められていることも事実であり，今後の技術的な発展が求められている。

4．自動フィードバックで活用できる言語情報

4.1　記述統計量

本節では，具体的に自動フィードバックにおいて活用できる言語情報について検討していく。技術的に自動フィードバックが比較的容易で，学習者自身にも分かりやすい言語項目として，総語数や異語数などが挙げられる。総語数は，ライティングの長さを表す指標であり，制限時間のあるライティング課題の場合に書き手の習熟度と高い相関を示すことが知られている（e.g., 小林・金丸，2012）[35)]。また，異語数は，語彙の豊富さと関わる指標である。そして，異語数を総語数で割った異語率を用いることで，ライティングにおいて語彙が繰り返し使われている度合いを数量化することができる。

$$\text{異語率} = \frac{\text{異語数}}{\text{総語数}}$$

異語率は，古くから使われている指標だが，テキストの総語数の影響を強く受ける，という欠点を持っている。（少なくとも理論上は）総語数を無限に増やしていけるのに対して，世の中に存在する単語の数（異語数）は有限であり，テキストが長くなるにつれて異語率が下がっていく傾向が見られる。従って，（総語数が大きく異なる）複数のテキストから求めた語彙多様性の値を比較する場合は，他の指標を用いることが望ましいとされている（Baayen, 2008）[36)]。

たとえば，総語数の影響を緩和するために提案された指標の一つとして，ギロー指数がある。この指標は，異語数を総語数の平方根で割ることで求めることができる。

$$\text{ギロー指数} = \frac{\text{異語数}}{\sqrt{\text{総語数}}}$$

語彙の多様性は，語彙研究や文体研究などの幅広い分野で活用されており，異語率やギロー指数以外にも様々な指標が提案されている（Baayen, 2001; Malvern, Richards, Chipere, & Duran, 2004）[37),38)]。

また，平均文長や平均単語長などの指標をライティング評価に用いることもある。これらの指標は，多くの単語からなる文，多くの文字からなる単語は難しい，という仮説に基づき，以下のように計算される。

$$\text{平均文長} = \frac{\text{総文数}}{\text{総語数}}$$

$$\text{平均単語長} = \frac{\text{総語数}}{\text{総文字数}}$$

4.2 品詞情報・構文情報

自然言語処理の技術を用いて，テキストに品詞や文法，構文などの情報が付与できる。まず，ライティングで用いられている個々の単語に品詞情報を自動付与し，品詞構成率を算出することで，ライティングの文体的特徴を概観することが可能になる。たとえば，学習者コーパス研究では，初級の学習者のライティングに名詞や代名詞が特徴的である一方，上級の学習者のライティングでは副詞や関係詞が効果的に用いられていることが知られている（e.g., Granger, 1998）[39)]。また，日本人英語学習者の場合は，日本語には存在しない冠詞の頻度が習熟度と関連していることもある。

なお，TreeTagger[注3)]やCLAWS[注4)]のような品詞情報自動付与ツールを用いると，テキスト中の時制や態などの文法情報も付与される。この情報を活用することで，初級者と上級者の間に存在する文法項目の使用傾向の差異を数量化することができる。

さらに，Stanford Parser[注5)]やCharniak Parser[注6)]などの構文解析器を用いると，名詞句や前置詞句の長さ，関係節の埋め込みの深さといった構文上の特徴を分析することもできる。学習者のライティングには文法的な誤りや不自然さが存在するため，母語話者による文章を解析した場合と比べて，構文解析の精度が低くなることもある。しかし，近年は，L2 Syntactic Complexity Analyzer[注7)]のような学習者の文章を構文解析するためのツールも開発されている（Lu, 2010）[40)]。

4.3 リーダビリティ

文章の難しさを測るための指標として，リーダビリティが利用可能である。様々なリーダビリティの指標が提案されているが，代表的なものとしては，Flesch Reading Ease（Kincaid, Fishburne, Rogers, & Chissom, 1975）[41)]が挙げられる。Flesch Reading Easeは，以下の公式で計算され，0～100の数字をとる。そして，数字が多いほど読みやすいことを表し，60～70が標準とされる。

Flesch Reading Ease

$$= 206.835 - 1.105\left(\frac{\text{総語数}}{\text{総文数}}\right) - 84.6\left(\frac{\text{総音節数}}{\text{総語数}}\right)$$

また，Flesch Reading Easeのバリエーションとして，Flesch–Kincaid Grade Levelがある。これは，以下の公式で計算され，計算結果は米国の学年に相当し，数字が大きいほど文章が難しいことを示す。因みに，6～10程度が望ましいと言われている。

Flesch-Kincaid Grade Level

$$= 0.39\left(\frac{\text{総語数}}{\text{総文数}}\right) + 11.8\left(\frac{\text{総語数}}{\text{総文数}}\right) - 15.59$$

リーダビリティは，複数の言語項目（総語数，総文数など）を一つの値に要約したものであるために学習者が結果を解釈しやすい，という利点を持っている。

4.4 キーワード

前項までに言及した言語項目は，ライティングの質を数量化するための指標である。それらの指標は，研究者にとって身近なものであるが，教師や学習者自身にとっては必ずしも分かりやすいものではない。そこで，フィードバックという観点では，実際にどのような語句を使うことができて，どのような語句を使うことができないのか，という具体的な情報が必要となる。そのような情報を得るには，コーパス言語学の分野で開発されたキーワード抽出の技法が有用である。

コーパス言語学におけるキーワード抽出は，二つのデータにおける全ての単語の使用頻度を比較し，どちらかに顕著に多く出現している単語を自動抽出する技術である。二つのデータにおける頻度の比較においては，カイ二乗値や対数尤度比のような検定統計量，オッズ比やクラメールのVのような効果量が用いられる。

たとえば，英語母語話者が書いたライティングと日本人英語学習者が書いたライティングを特徴づける単語を抽出することで，「母語話者らしい英語」や「日本人らしい英語」を分析することが可能になる。そして，自動フィードバックにおいては，学習者のライティングとより質の高い（より高い習熟度を持つ書き手による）ライティングと比較し，学習者が統計的に過少使用している単語に注目することで，「よりよい英語を書くために習

得する必要のある表現」が分かる。また，学習者が統計的に過剰使用している単語に注目することで，学習者が「繰り返し用いている表現」（多くの場合は，他の表現に言い換え可能な表現）が明らかになる。

4.5 *n*-gram

ライティングにおける文体的特徴を分析する場合は，文章における *n* 個の要素の連鎖である *n*-gram を抽出する。まず，習熟度の高い学習者のライティングとそうでない学習者のライティングにおける高頻度な単語 *n*-gram を抽出し，両者を比較することで，「よりよい英語を書くために習得する必要のある表現」や学習者が「繰り返し用いている表現」を具体的に把握することができるようになる。また，品詞の連鎖からなる品詞 *n*-gram を抽出することで，ライティングのトピックの影響を軽減し，より抽象化・一般化されたレベルで文体的特徴を分析することが可能になる。単語 *n*-gram や品詞 *n*-gram は，ライティングの自動採点でも頻繁に用いられる言語項目であるため（小林，2017）[42]，自動採点システムと連携したフィードバックシステムでも活用することは理にかなっている。

5．まとめ

本稿では，これまでの自動採点やフィードバックに関する理論と技術を概観し，自動フィードバックで活用できる言語項目について議論してきた。今後，実際のシステムを開発するにあたっては，前段で議論したように，フィードバックの粒度について検討する必要がある。たとえば，システムが扱うことのできる全ての項目フィードバックを与えるか，任意に設定した言語項目のみに限定するか，を選択できるようにする（フィードバック対象の選択）などが考えられる。また，文法的誤り訂正に関しては，誤りを含む箇所を指摘するのみか，誤りの種類をヒントとして示すか，正解候補を直接示すかなどを調節できることが望ましい（フィードバックの明示性の選択）。そして，教員や学習者にフィードバックを提供する際，どれぐらい詳細な情報を提示するか，どのような形式の図表や文章を用いるか，などについては，診断的フィードバックに関する実証的研究（e.g., Roberts & Gierl, 2010）[43]の知見を活用すべきである。

謝辞

本研究は JSPS 科研費 17K13511 および JSPS 科研費 16K16885 の助成を受けたものである。

注

注１）たとえば，以下のプレスリリースを参照。
https://www.eiken.or.jp/eiken/info/2018/pdf/20181017_pressrelease_aisaiten.pdf

注２）文法誤り検出・訂正の研究動向については，永田（2017）[44]に詳しい。

注３）http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

注４）http://ucrel.lancs.ac.uk/claws/

注５）https://nlp.stanford.edu/software/lex-parser.shtml

注６）http://bllip.cs.brown.edu/resources.shtml#software

注７）http://www.personal.psu.edu/xxl13/downloads/l2sca.html

参考文献

１）大関浩美（編）『フィードバック研究への招待―第二言語習得とフィードバック』，くろしお出版，2015.

２）Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. New York: Routledge, 2003.

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

４）Ueno, M., Web based computerized testing system for distance education. *Educational Technology Research,* 28, 2005, 59-69.

５）Shin, D., Min, H., Park, S., Jung, C. K., Joo, H., & Kim, M., Validation research for developing and applying the automated scoring program for the speaking section of the NEAT. *The Abstract Book of the 35th Annual Language Testing Research Colloquium*, p. 44, 2013.

６）Page, E. B., Project Essay Grade: PEG. In Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. New York: Routledge, 2003, 43-54.

７）Elliot, S., IntelliMetric: From here to validity. In Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. New York: Routledge, 2003, 71-86.

８）Landauer, T. K., Laham, D., & Foltz, P. W., Automated scoring and annotation of essays with

the Intelligent Essay Assessor. In Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. New York: Routledge, 2003, 87-112.
9) Burstein, J., The e-rater® scoring engine: Automated essay scoring with natural language processing. In Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. New York: Routledge, 2003, 113-121.
10) Liao, H., Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal*, 70(3), 2016, 308-319.
11) Williamson, D. M., Developing warrants for automated scoring of essays. In Shermis, M., & Burstein, J. (Eds.), *Handbook of automated essay evaluation*. New York: Routledge, 2013, 153-180.
12) Bejar, I. I., Williamson, D. M., & Mislevy, R. J., Human scoring. In Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.), *Automated scoring of complex tasks in computer-based testing*. Hillsdale: Lawrence Erlbaum Associates, 2006, 49-81.
13) Ling, G., Mollaun, P., & Xi, X., A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing*, 31(4), 2014, 479-499.
14) Ericsson, P. F., & Haswell, R. (Eds.), *Machine scoring of student essays*. Logan: Utah State University Press, 2006.
15) Higgins, D., Ramineni, C., & Zechner, K., Learner corpora and automated scoring. In Sylviane, G., Gilquin, G., & Meunier, F. (Eds.), *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press, 2015, 587-604.
16) Attali, Y., Validity and reliability of automated essay scoring. In Shermis, M., & Burstein, J. (Eds.), *Handbook of automated essay evaluation*. New York: Routledge, 2013, 181-198.
17) Kobayashi, Y., & Abe, M., Automated scoring of L 2 spoken English with random forests. *Journal of Pan-Pacific Association of Applied Linguistics*, 20(1), 2016, 55-73.
18) Xi, X., Validity and the automated scoring of performance tests. In Fulcher, G., & Davidson, F. (Eds.), *The Routledge handbook of language testing*. New York: Routledge, 2012, 438-451.
19) Larkey, L. S., & Croft, W. B., A text categorization approach to automated essay grading. In Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. New York: Routledge, 2003, 55-70.
20) Biber, D., Nekrasova, T., & Horn, B., The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *TOEFL-iBT Research Report*, 14, 2011, 1-99.
21) Truscott, J., The Case Against Grammar Correction in L2 Writing Classes. *Language Learning*, 46, 1996, 327-369.
22) Ferris, D., The case for grammar correction in L2 writing classes: a response to Truscott (1996). *Journal of Second Language Writing*, 8(1), 1999, 1-11.
23) Ellis, R., A typology of written corrective feedback types. *ELT Journal*, 63(2), 2009, 97-107.
24) Han, Y., Written corrective feedback from an ecological perspective: The interaction between the context and individual learners. *System*, in press.
25) Amano, S., Students' choices of types of written teacher feedback in EFL writing instruction. *Journal of the Chubu English Language Education Society*, 2018, 103-110.
26) Han, Y., & Hyland, F., Academic emotions in written corrective feedback situations. *Journal of English for Academic Purposes*, 38, 2018, 1-13.
27) Leacock, C., Chodorow, M., & Tetreault, J., Automatic grammar- and spell-checking for language learners. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 2015, 567-587.
28) 辻井潤一,「合理主義と経験主義のはざまで―内的な処理の計算モデル」, 人工知能学会誌, 27(3), 2012, 273-283.
29) Dikli, S., The nature of automated essay scoring feedback. *CALICO Journal*, 28(1), 2010, 99-134.
30) Ranalli, J., Automated written corrective feedback: how well can students make use of it? *Computer Assisted Language Learning*, 2018, 1-22.
31) 永田亮,「構文解析を必要としない主語動詞一致誤り検出手法」, 電子情報通信学会論文誌, D, 情報・システム, 96(5), 2014, 1346-1355.
32) Sakaguchi, K., Hayashibe, Y., Kondo, S., Kanashiro, L., Mizumoto, T., Komachi, M., & Natsumoto, Y., NAIST at the HOO 2012 Shared Task. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 2012, 281-288.

33）Rozovskaya, A., Chang, K.-W., Sammons, M., & Roth, D., The University of Illinois system in the CoNLL-2013 shared task. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, 2013, 13-19.
34）Tajiri, T., Komachi, M., & Matsumoto, Y., Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 2012, 198-202.
35）小林雄一郎・金丸敏幸,「パターン認識を用いた課題英作文の自動評価の試み」, 電子情報通信学会技術研究報告, 112（103）, 2012, 37-42.
36）Baayen, R. H., *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press, 2008.
37）Baayen, R. H., *Word frequency distribution*. Dordrecht: Kluwer Academic Publishers, 2001.
38）Malvern, D. D., Richards, B. J., Chipere, N., & Duran, P., *Lexical diversity and language development: Quantification and assessment*. Hampshire: Palgrave Macmillan, 2004.
39）Granger, S.（Ed.）, *Learner English on computer*. London: Longman, 1998.
40）Lu, X., Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15⑷, 2010, 474-496.
41）Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S., *Derivation of new readability formulas*（*Automated Readability Index, Fog Count and Flesch Reading Ease Formula*）*for navy enlisted personnel*. Millington: Naval Technical Training, U. S. Naval Air Station, Memphis, Tennessee, 1975.
42）小林雄一郎,「英語の自動作文評価」李在鎬（編）『文章を科学する』, ひつじ書房, 2017, 158-174.
43）Roberts, M. R., & Gierl, M. J., Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, 29⑶, 2010, 25-38.
44）永田亮,『語学学習支援のための言語処理』コロナ社, 2017.

（H 31. 2 . 9受理）