英語初級学習者のパラグラフ・ライティング自動評価システム開発の試み Part3

―生成 AI を用いた英文の質を評価するシステムの分析

An Attempt to Develop an Automated Evaluation System for Paragraph Writing of Pre-intermediate Learners of English: Part 3
―Analysis of a System for Evaluating the Quality of English Text Using Generative AI

MITA Kaoru
三　田　　薫
国際学部国際学科教授
SHIMODA Atsuko
霜　田　敦　子
国際学部国際学科非常勤講師

抄録：

短期大学 1 年生向け英語必修科目で実施してきたイティングテストのデータを用いた自動評価システムを開発した．生成 AI の大規模言語モデル（GPT-4o）に過去の学生の英文エッセイ 1611 件と教師評価を学習させ，自動採点システムを作成し，前年度まで開発していた人工知能の機械学習の「教師あり学習」モデルとその精度を比較検証した．自動評価システムは英文の「内容の質」のみを評価対象とし，4 段階の評価を行う．生成 AI モデルと機械学習モデルの両方で 2024 年度の英語必修科目のライティングテストの英文 129 件を評価したところ，生成 AI モデル（GPT-4o）では，教師評価との一致率が 83.7% であった．一方，機械学習モデルでは教師評価との一致率が 59.7% であった．生成 AI モデルによる自動評価システムを授業内で学生に使用させた上でアンケート調査を行い，それをテキストマイニングで分析した．

Abstract：

We developed an automated evaluation system using data from writing tests conducted in a compulsory English course for first-year junior college students. A generative AI model (GPT-4o) was trained on 1,611 English essays written by past students, along with their

corresponding teacher evaluations, to create an automated evaluation system. Its performance was then compared to a supervised machine learning model developed in previous years. The automated evaluation system assessed only the “quality of content” of the English essays, using a 4-level scale. Both, the generative AI, and machine learning models, were used to evaluate 129 English essays from the 2024 academic year’s compulsory English writing test. The generative AI model（GPT-4o）achieved an agreement rate of 83.7% with teacher evaluation. In contrast, the machine learning model revealed an agreement rate of 59.7% with teacher evaluations. Following its implementation in the classroom, we conducted a survey among students to gather feedback and analyzed the results using text mining.

キーワード：自動評価システム，人工知能，機械学習，生成 AI，ChatGPT，一致率，内容の質，第 2 言語ライティング，相関分析，テキストマイニング

Keywords：Automated Evaluation System（AES），Artificial Intelligence，Machine Learning，Generative AI，ChatGPT，Agreement Rate，Quality of Content，Second Language Writing，Correlation Analysis，Text Mining

1. はじめに

筆者らは短期大学で 1 年次英語必修科目（Integrated English）[1)] において，過去数年間にわたり特定のテーマで年 3 回ライティングテストを行い，そのデータを分析してきた（三田・霜田，2020, 2021a, 2021b, 2022a, 2022b, 2023a, 2024a）．2021 年度には，調査データの一部について人工知能の「機械学習」を用いた自動評価システムを開発する試みを開始した．このシステム開発は，授業担当者の英文ライティングの評価に関わる作業負荷の削減と学習者の自律的学習の促進を目的としている．この自動評価システムを学期末のライティングテスト準備のために学生に使用させた結果，英文の「内容の質」向上に大いに貢献していたことがわかった（三田・霜田，2023b）．しかしながら，我々の取り組みのような授業単位での小規模のデータ収集では，機械学習によるシステム開発のためのデータ量は限られており，おのずと教師評価との一致率の向上には限界があった．こうした中，2022 年 11 月に AI 言語モデル ChatGPT が公開された．その後，生成 AI の言語教育への応用や実践研究は急激に進み，現在もさらなる発展が期待されている．本稿ではライティングの自動評価システムとして新たに開発した生成 AI モデルを，これまでの機械学習モデルと比較検証し，授業で学生に使用させた結果について報告する．

第 2 節では英語教育の現場における生成 AI を用いた自動評価システムについての先行研究を紹介し，第 3 節でリサーチクエスチョン，第 4 節で調査方法，第 5 節で調査結果を述べ，第 6 節で考察を行い，第 7 節でまとめる．

2. 先行研究

ライティングの自動評価（Automated Evaluation System: AES）は，人による評価に伴う時間と労力を軽減し評価結果の信頼性の問題といったリスクを回避できるため，50 年以上前に始まって以来研究者や教育者の間で注目を集めてきた（cf. 石井・近藤, 2020）．特に 1990 年代のコンピュータの進歩により，より信頼性のある自動評価システムが開発され，その代表が 1998 年に Educational Testing Service（ETS）により開発された e-rater である．e-rater は当時最先端の自然言語処理技術を駆使し，高い信頼性と妥当性を持ち（Attali & Burstein, 2004）TOEIC, TOEFL といった資格試験において使用されている．e-rater には評価レベルの判定だけでなく，エラーに対する自動フィードバックを提供する Criterion が搭載されており，教育現場で広く活用されている（小林, 2017）．Li et al.（2015）は，Criterion がライティング指導とパフォーマンスにどのような影響を与えたかを調査し，その添削フィードバックが，下書きから最終稿までの精度の向上に役立つことを示唆している．また，Almusharraf and Alotaibi（2022）は Grammarly（人工知能と自然言語処理を用いたデジタルライティングツール）の自動評価システムによる評価と人間の評価者による評価を比較し，EFL 学習者エッセイ評価について，人間の評価と Grammarly の結果の相関は中程度であること，人間の評価者の方が高いスコアを出していることを示している．一方，Zribi and Smaoui（2021）はチュニジアの中レベルの英語力の大学生エッセイについて，Paper Rater という自動評価システムの評価が人間の評価者の評価を大幅に上回っていたと報告している．

このように自動評価研究が盛んに行われてきた中で，2022 年 11 月に OpenAI の GPT-3.5 が公開され，ライティング評価研究にも大きな影響をもたらすこととなった．ChatGPT はコンピュータプログラムを使うことができる専門家だけでなく，誰もが言語生成 AI ツールを使用することを可能にした．すでに教育現場においても生成 AI の使用は浸透してきており，第二言語学習とその研究においても，人間と生成 AI が共存する新しい時代に入っている．

生成 AI の使用が教育分野に与える影響については，様々な議論がなされている．Nature Editorial（2023）は，大きな懸念として，出版者や研究者・学生による無責任な使用に言及している．Nature 誌においては，生成 AI による研究論文の著者は信頼される著者と認めないこと，またその使用を文書内のどこかに明記することという 2 つの原則を規定している．EduKitchen（2023）と Chomsky の対談動画で，Chomsky は ChatGPT を使用することは「言語を理解するということにおいて価値がない（… no value with regard to understanding about language … 5:36）」と述べ，ChatGPT を「ハイテク剽窃（high-tech plagiarism）」と批判している．オーストラリア大学では ChatGPT を使って作成されたエッセイが増加していることを問題視し，テストにおいて紙と鉛筆で書く形式にするとともに評価基準も変更したことがニュースで報じられた（Goodall, 2023）．Kohnke et al.（2023）は学習における ChatGPT の教育支援ツールとしての利便性を認めつつ，その欠点とリスクについて指摘している．それは，不正利用といった倫理的問題，その回答が適切に引用されていないソースの言い換え（つまり盗用）であること，さらに回答に誤りがある可能性，データベースのテキストの大半は英語コーパスから抽出されているた

め，文化的に中立ではない点である．こうしたリスクや欠点を克服するためには，教師と生徒の高度なデジタルコンピテンシーが必要であると主張している．

一方で，今後も発展する生成 AI の恩恵を受けるべきであるとする意見も多い．ジャーナリズムとメディア研究者である著者と ChatGPT による共著という形で，Pavlik（2023）は，ChatGPT の回答を引用しその精度を読者に判断させるという試みを行っている．ChatGPT の回答には一部の知識に範囲や深さの限界があるものの，全般的に素晴らしいレベルと範囲の知識を持つことを示している．このことにより，ChatGPT はジャーナリストやメディアの専門家を支援するためのツールとして，その仕事の質と効率を向上させる可能性を示唆している．

生成 AI を外国語教育に使用することには利点と欠点がある．Essel（2023）は，ChatGPT の教育への導入の利点として，生徒の質問に即座に，正確に，個人にあった応答を提供し，小論文の評価を自動化し，より個別化された学習体験を提供することを挙げている．一方デメリットとして，生徒の批判的思考能力の低下を招く可能性や回答に誤りがあることもあると指摘している．しかし，ChatGPT は生徒の学習体験を向上させ，教育をより効率的かつ効果的にすることができる貴重なツールであるため，教育者や研究者は既存の教育システムに統合させながら，欠点を最小限に抑えつつ，その可能性を最大限に引き出す方法を見つけることが最も重要であると主張している．Teng（2024）は，L2 ライティングのプロセスに ChatGPT を活用し，ChatGPT によるフィードバックが大学生のライティングに与える影響を調査し，ChatGPT のような AI ツールによるライティング支援がライティングに有意なプラスの効果をもたらしたことを実証している．しかし ChatGPT は教師が提供するような人間味のあるフィードバックができないという学生の不満もあり，教師の役割を完全に代替することはできないため，教師はフィードバックを補足することが重要であり，ChatGPT は「敵ではなく，仲間（a companion，not enemies）」として機能すると主張している．

ライティングの自動評価に ChatGPT を使用する研究も現れ始めている．Mizumoto and Eguchi（2023）は「GPT は様々な言語生成タスクでは優れた結果を示しているが，AES ではまだ利用されていない（p. 6）」（筆者翻訳）とし，GPT-3.5 を活用して ETS の非ネイティブ筆記英語コーパス（TOEFL11）に含まれる小論文の自動評価を行い，その信頼性と精度の評価を試みている．その結果，GPT は一定の精度と信頼性を有し，人間による評価を裏付けることがわかった．特に，言語的特徴（語彙の多様性，構文の複雑さ，細かい構文特徴，動詞と項の関係（統語構造），テキストの結束）を評価に組み入れた結果の精度が向上した．生成 AI が自動評価ツールとしてライティング評価に効果的に活用できること，また研究と実践の両方において，ライティング評価とフィードバックの方法に革命をもたらす可能性があることを示唆している．ただし，GPT を用いた自動評価システムは一定の精度を達成することができるが，それでも人間の評価者との完全な一致を達成するまでには至らないため，人間による評価と併用されるべきであると，あくまでも補助的なツールであることを強調している．

ChatGPT を使ってライティングの言語的正確性を評価した Pfau et al.（2023）では，Grammarly のような既存の AI プログラムは AI と人間のやり取りができないが，ChatGPT は具体的で複雑

なニーズに合わせてカスタマイズできるというこれまでにない特徴を指摘している．その研究結果では，GPT-4 によるエラー検出と人間による評価に強い相関関係があることを示した．ただし，初級学習者のエッセイではエラーの見落としが多く，人間による評価との相関が低く，また結果の一貫性に問題があるといった ChatGPT の限界を強調し，今後の研究課題としている．Uchida（2024）は，ChatGPT が学習者のライティングをどの程度正確に評価できるか調査した結果，人間が評価したライティングの総合得点と GPT-3.5 の得点との相関は 0.801，GPT-4 では 0.888 であった．この結果は，生成 AI のスコアが一定の信頼性を持つことを示している．

生成 AI と機械学習の評価精度を比較した研究もある．Mizumoto et al.（2024）は，ChatGPT による言語的な正確さの評価を人間の評価者と Grammarly と比較した結果，ChatGPT による評価と人間による評価に強い相関関係があり，Grammarly との相関関係を上回ったことが明らかになった．この結果は，ChatGPT による L2 ライティングの自動評価の可能性を肯定するものである．

また，議論型エッセイ（argumentative essay）の構成の評価に焦点を当てた Baffour et al.（2024）では，GPT-4 で評価した議論型エッセイの全体的評価は人間と同等の精度があったが，一方で，構造的な部分（主張，反論，裏付けとなる証拠など）の分析的評価の精度は低かったことを明らかにしている．

このように ChatGPT によるライティングの自動評価の精度を調査する研究は始まったばかりであるが，いろいろな角度からアプローチを試みる研究が急速に増えつつある．筆者らは，これまで継続してきた「内容の質（quality of contents）」を測る自動評価システムを，機械学習から ChatGPT に変えて開発することを試みる．本研究の目的は，ChatGPT による内容の質評価と人間による内容の質評価との精度の相関を調査し，開発された自動評価システムが学習者のライティング力向上に有益であるかを検証することである．

3. リサーチクエスチョン

（1）ChatGPT と機械学習による「内容の質」を測る自動評価システムはどの程度の精度があるか
（2）自動評価システムを利用した学生のアンケートから示唆されるものは何か

4. 調査方法

4.1. 自動評価システムの概要

これまでに開発した Model A から Model C までは，Amazon 社がクラウドコンピューティングを介して提供する Amazon Machine Learning の機械学習機能を活用し，自動評価システムの開発に関する実験的研究を実施した[2)]．自動評価システムの分野の先行研究において導入されている「特徴量」は，本研究では導入していない．また 1 回のテストで大量のデータが収集できる検定試験や入学試験に比べて，入力できるデータ件数も圧倒的に不足している．データ数の不足を克服するため，1）ライティングのテーマを 1 つに限定する，2）毎回同じテーマでテストを実施することにより，データの積み上げを可能にする（学生には同じテーマであっても毎回異なる対象を選択することを義務付けている），3）テストで使用する表現も可能な限り限定する，4）

エッセイの構成や，使うべきディスコースマーカーを指定する，5）文法やスペリングのエラーは評価の対象とせず，「内容の質」に限定して 4 段階の評価を行うといった方策を導入している．このアプローチを用いて，2020 年度，2021 年度，2022 年度の授業内でライティングテストを年 3 回実施し，さらに 2023 年度 1 回目のデータを加え，それらのデータに教師の評価結果を「正解」として加えて機械学習の「教師あり学習」を実施した．

ライティングテストの主題を一貫して同一にする背景には，データ量の増加という実用的な動機のみならず，学生が特定のテーマに反復的に取り組むことで，パラグラフ・ライティングの構造や内容の深化に関する学びを習得するためという意図がある．そのため，このライティングテストは総括的評価（学習の最後に，生徒の学力の達成度を確認するために行う評価）ではなく，形成的評価（学生がライティングを改善できるようにサポートするための評価）に重点を置いている．

第 1 期モデル（Model A）は 2022 年 9 月授業開始前に開発され，第 2 期モデル（Model B）は 2022 年 1 月最終授業前に開発されている．第 3 期モデル（Model C）は，2023 年後期授業開始前に開発された．第 4 期モデル（Model D）は，それまでの機械学習とは異なり，生成 AI を用いた自動評価システムである．Model A，B，C，D それぞれが使用した英文エッセイは以下の表 1 の通りである．前期始め，前期末，後期末の 3 回分のライティングテストの英文をデータベースとして使用している．

表 1　自動評価システムの各モデルで使用したデータベースの英文エッセイ

Model A	2021 年度 3 回分
Model B	2021 年度 3 回分，2020 年度 3 回分
Model C	2020 年度 3 回分，2021 年度 3 回分，2022 年度 3 回分，2023 年度 1 回分
Model D	2020 年度 3 回分，2021 年度 3 回分，2022 年度 3 回分，2023 年度 2 回分

各モデルの開発では，それ以前のモデルにデータを追加するのではなく，その時点までに揃っているデータをまとめて機械学習に読み込ませている．Model D 開発には，表 2 の計 1611 件の英文をデータベースとして使用した．Model C 作成で利用したデータは 1506 件，Model D 作成で使用したデータは 1611 件である．

表 2　Model D 開発に用いたデータベースの英文エッセイ

英文回収時期	英文件数
2020 年度 5 月 7 月 1 月	482
2021 年度 4 月 7 月 1 月	509
2022 年度 4 月 7 月 1 月	394
2023 年度 4 月 7 月	226
合計	1611

Model D 開発に先立ち，英文エッセイの改行の削除，句読点の統一などの「前処理」を行った．その後，OpenAI 指定の学習用データ作成のための記述ルールに基づいてデータを作成した．

前回の Model C の作成の際は，Amazon Machine Learning を利用してモデル開発を行った．機械学習には検証データが必要なため，Model C 開発では 70％を学習用に 30％を検証用に使用した．その研究との連続性を持たせるため，Model D 開発にあたっては，Model C 同様，70％を学習用に 30％を検証用に使用したモデルを別途作成した．これを 100％エッセイを学習したモデルと比較し，より性能の高いモデルが選択できるようにした[3)]．

表 3　Model D 用に開発した 2 種類の学習用モデル

	学習用エッセイ数	検証用エッセイ数
英文 100％を学習したモデル	1611 件	0 件
英文 70％を学習に 30％を検証に使用したモデル	1128 件	483 件

Model D 開発に用いた ChatGPT では，Fine-tuning の 2 種類のパラメータ（TopP と Temperature）を用いた．これらは，ChatGPT における多様性・ランダム性，不規則性を調整することができるパラメータである．ChatGPT のような LLM モデルは特定の単語の前後に一緒に使用させる可能性の高い単語を推測しながら文字列を作成している．出現確率が高いものはより無難な回答となり，出現頻度が低い単語の配列はより多様な表現となるが，一方で誤りが増加する傾向があるため，これらのパラメータを使用することによってその調整を行うことが可能となる．

Model D 開発にあたって，当初は GPT-3.5 を用いて調査を行った．開発時点（2024 年 8 月 6 日）では，GPT-4o による Fine-tuning は「Tier 4・5」として区分された組織のみが利用可能であったため，やむなく GPT-3.5 で Fine-tuning を行うこととなったという経緯がある．しかしその後この区分が解除されたため，GPT-4o で同一のエッセイデータを用いて，改めて調査することとなった．

Fine-tuning を用いた事前調整を行わない場合，入力データに対して ChatGPT では適切な判断が行われないことがある．ChatGPT は膨大なインターネット上のテキスト情報を学習し，単語間の関連性を推測しながら文章を生成するモデルであるため，Fine-tuning 無しのデフォルトの状態で使用すると，2024 年 8 月時点においては以下のような問題が発生している．

（1）デフォルトではエッセイを評価する際の数値計算が苦手
（2）デフォルトでは明確なエッセイ評価基準がない
（3）デフォルトでは模範となる評価結果を学習していない

仮に（2）の「明確なエッセイ評価基準」をプロンプトによって調整しても，（1）の評価に必要となる正確な数値計算が行えないため，エッセイの多角的な評価が難しい状況であった．

Fine-tuning を使用することで，(3)「模範となる評価結果」を大量に学習させることが可能となり，結果として (2)「明確なエッセイ評価基準」をモデル化させることにより，より人間の評価に近い結果が出力できるようになった．

Model D の開発にあたっては，何度も試作を繰り返した．ChatGPT にプロンプトによってエッセイを評価させた際には，入力の度に異なる結果を出力したり，詳細な指示を含むプロンプトを用意しても，そうした指示が守られなかったりといったことが頻発した．その後 Fine-tuning を使用することによって，詳細なプロンプトを用意せずとも，「内容の質」をダイレクトに ChatGPT の評価に反映させることができるようになった．「内容の質」評価という人間の評価が内包された今回のエッセイ評価は，プロンプトによる言語指示では伝えきれない文章から感じ取れる印象や躍動感，感情などの評価をコンピュータに学ばせることができる可能性を示唆している．

4. 2. 機械学習入力用英文のトピック

機械学習の入力データとなる 2020, 2021, 2022 年度のライティングテストは，すべてオンラインで実施した．具体的には学習管理システム（LMS）の manaba ver.2.95（朝日ネット）を用いて受験させた．ライティングテストの所要時間は 2022 年度までは 15 分間で，入力画面には，ライティングの入力用スペースだけではなく，受験者個人のブレインストーミング内容を記録するためのスペースも設けた．2023 年度 1 回目以降のライティングテストは，あらかじめ配布した用紙に 20 分間手書きさせた上で，試験終了後に，manaba の画面に手書きした英文を打ち込ませた．試験時間および試験方法の変更の理由は，タイピングに不慣れなことが原因で本来の実力が発揮できないことを解消するためである．

ライティングテストのトピックは「好きな場所」（意見文）である．意見文（Opinion Essay）は英語検定試験において頻繁に出題されるジャンルであるため，実用面からもそうした試験に準拠したトピックとした．エッセイの指示文は以下の通りである．

> ライティングテストトピック「好きな場所」
> 自分の行ってみたいところを決め，その場所と，行きたい理由を 3 つ書いて下さい．海外でも国内でも結構です．以下の表現で書いてください．
>
> The place I would like to visit the most is（　　）. There are three reasons.

年 3 回のライティングテストとも同じテーマを用いている．ただし学生には必ず 3 回とも別な場所を選んで書くよう指示し，同じ場所が選ばれている英文には一切加点されないことを伝えている．

4. 3. 学生英文の評価基準

ChatGPT モデル（Model D）においても機械学習の「教師あり学習」（Model C）と同じく，「内容の質」に関する以下の４つの評価基準 Level 1 から Level 4 で評価した．ただし，未完のもの，理由が３つ無いもの，トピックが違うもの，単語羅列で意味が伝わらないものなどは，最低限の修正を施し Level 1 以上の英文にリライトした．「Detail 文」とは，詳しい説明や具体例を挙げて，主張に説得力を与える文のことである．「内容の質」については，Wiseman（2012）のライティング・ルーブリックにおける Topic Development という評価項目を応用している．

Level 1：Detail 文が無いもの
Level 2：Detail 文が少しあるが内容が限定的なもの
Level 3：Detail 文が複数ありトピックが発展し内容が深まったと考えられるもの
Level 4：Level 3 の英文の中で特に優れているもの

以下は各評価のサンプル英文である．「内容の質」の評価であるため，文法やスペリングのエラーが含まれていても，自動評価システムでは評価の対象としていない．以下は１から４の各レベルの英文サンプルである．

① Detail 文が無いもの「Level 1」例
The place I want to visit most is Okinawa. There are three reasons. First, I like hot place. Second, beach is beautiful and rich in nature. And finally, I've never eaten Okinawa food, and I want to try them.

② Detail 文が少しあるが内容が限定的なもの「Level 2」例
The place I want to visit most is Hawaii. There are three reasons. I want to eat delicious food in Hawaii. For example, I like garlic shrimp. I want to surf because my father is doing it. I want to go to the beach in the evening because the setting sun is beautiful.

③ Detail 文が複数ありトピックが発展し内容が深まったと考えられるもの「Level 3」例
The place I would like to visit most is Korea. There are three reasons. First, I want to go to different cities and go shopping because I've been watching Korean dramas every day lately. Therefore, the reason I want to go to Korea is probably influenced by Korean dramas. Second, I want to buy Korean cosmetics. They are so good for my skin. So, I often use Korean cosmetics. Finally, there are many delicious foods in Korea. I especially like spicy food. So, I want to eat a lot of spicy food when I go to Korea. For these reasons, I would like to visit Korea.

④ 「Level 3」の中で特に優れているもの「Level 4」例

The place I would like to visit most is Australia. There are three reasons. First, I love animals. Australia is a great place to get close to rare animals, such as koalas and kangaroos. My parents showed me a picture of them holding a koala and a crocodile when they traveled to Australia a long time ago, so I want to hold a koala or a crocodile too. Second, I want to visit the many World Heritage sites, for example, Ayers Rock, the Great Barrier Reef, and the Opera House. In particular, I want to see an orchestra concert at the opera house because Australia is famous for classical music. Finally, it is summer in Australia when it is winter in Japan. I don't like the cold. Therefore, I would like to spend time in Australia during the winter season in Japan. Moreover, the time difference between Australia and Japan is about an hour or two. I've never traveled overseas before, so I'm glad that there isn't much of a time difference. For these reasons, I would like to visit Australia.

4. 4. 自動評価システムの表示画面

今回開発した自動評価システム（Model D）の表示画面は以下の図 1，図 2 の通りである．図 1 は英文入力前の画面，図 2 は英文入力・自動評価後の画面である．

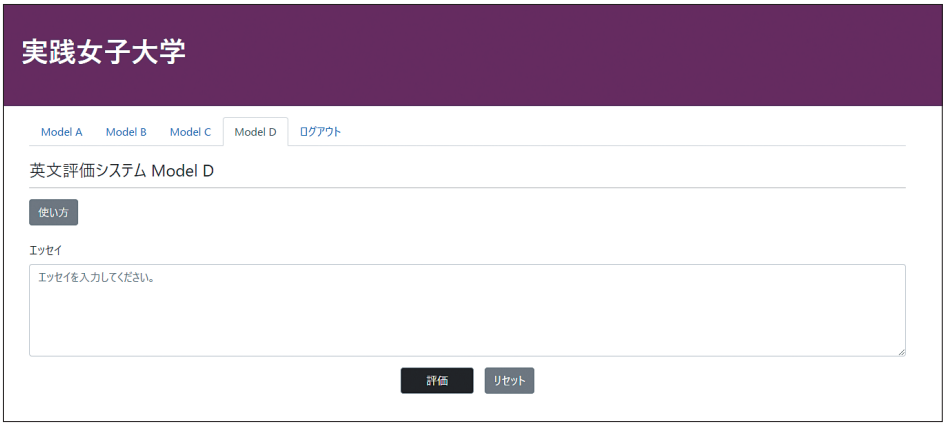

図 1　英文評価システムの英文入力前画面

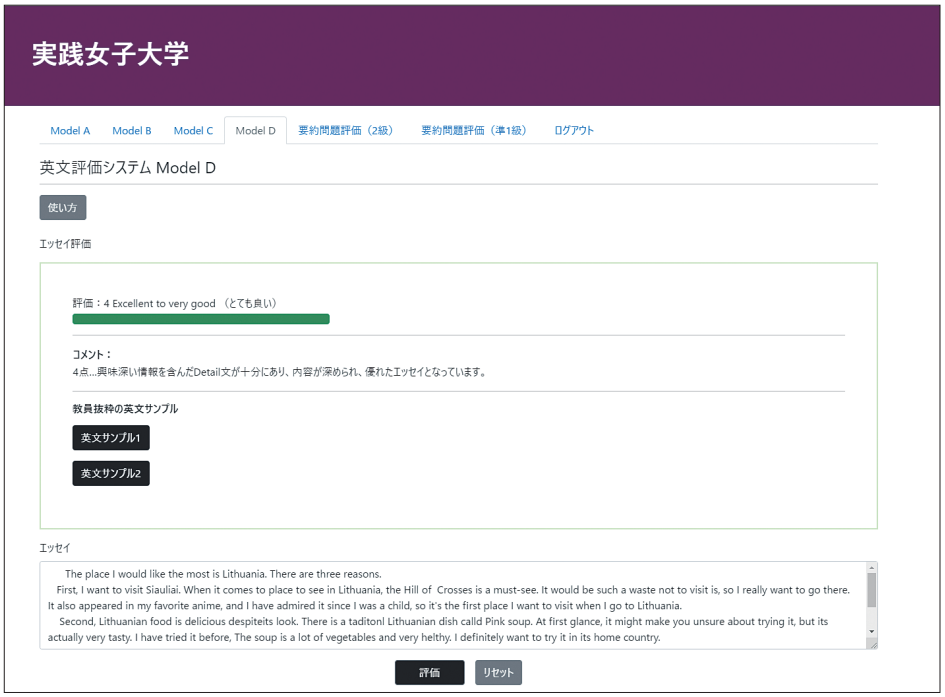

図２　英文評価システムの英文入力・自動評価後画面

表３は，英文の評価と同時に表示されるレベル別のフィードバック表現である．

表３　自動評価システムに表示されるフィードバック表現

Level	自動評価システムのフィードバック表現
1	評価：1 Very poor（良くない） コメント：1 点…Detail 文がありません．理由の詳細や具体例をそれぞれの理由に付けてみましょう．
2	評価：2 Fair to poor（あまり良くない） コメント：2 点…Detail 文が平凡です．読んだ人に印象が残るような情報（例えば自分の経験や皆が知らないような情報など）を入れてみましょう．
3	評価：3 Good to Average（良い） コメント：3 点…興味深い情報を含んだ Detail 文が複数あり，内容が深められています．
4	評価：4 Excellent to very good（とても良い） コメント：4 点…興味深い情報を含んだ Detail 文が十分にあり，内容が深められ，優れたエッセイとなっています．

第２期モデル（Model B）開発の際，自動評価システムで出力される評価得点と同時に表示される自動フィードバックとして，過年度の学生が同じトピックで作成した英文を，英文サンプル（典型事例）（Sadler, 1987: 岩田, 2020: 丹原他, 2020）として提示する機能を追加した（図３）．Model A，B どちらの出力画面にも付加し（三田・霜田, 2023b），その後開発された Model C,

D にも継承された．これにより自動評価システム Model A, B, C, D は，教師の自動評価ツールとしてだけではなく，学生の形成的評価につながるツールとしても機能する可能性を持つことになった．

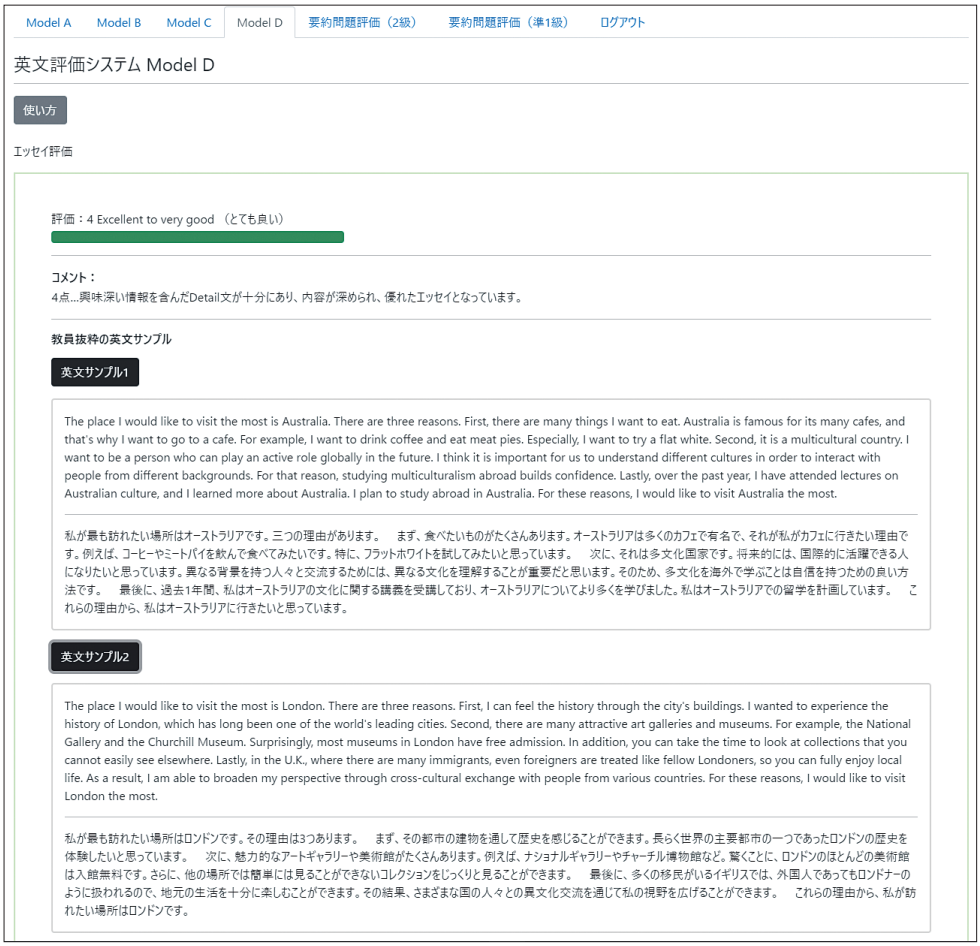

図 3　英文評価システムのモデル英文画面

4. 5. 分析方法

自動評価システム Model C，D について，以下の 2 つの分析を行った．

（1）教師評価と，機械学習システムおよび ChatGPT システムによる評価との一致率
（2）学生アンケートの自由記述のテキストマイニング分析

5. 調査結果

5. 1. 自動評価と教師評価の一致率と相関係数

2024 年 9 月の後期授業開始時に学生に Model D を使用させた．学生の英語レベルは CEFR

A2 中心である[4]．学生には 7 月に実施したライティングテストで自ら作成した英文エッセイを自動評価システムで評価させた[5]．表 4 は，学生英文 129 件を Model C，Model D（GPT-3.5），Model D（GPT-4o）で自動評価した結果と教師評価の一致率である．

表 4　2024 年 7 月の学生英文 129 件についての自動評価と教師評価の一致率

	一致英文件数	不一致英文数	英文総件数	一致率（%）
Model C（機械学習）と教師評価	77	52	129	59.7%
Model D（GPT-3.5）と教師評価	90	39	129	69.8%
Model D（GPT-4o）と教師評価	108	21	129	83.7%

欠席者 10 名と 0 評価 3 名を除く 129 名

表 5 は，Level 1 から Level 4 の 4 段階評価の自動評価と教師評価の平均と標準偏差である．

表 5　自動評価と教師評価の平均と標準偏差（N=129）

	M	*SD*
Model C（機械学習）	3.698	.5245
Model D（GPT-3.5）	3.628	.6740
Model D（GPT-4o）	3.550	.6245
教師評価	3.450	.6369

表 6 は，4 段階評価の自動評価と教師評価の相関係数である．相関係数は 1%水準で有意である．4 段階評価の Model C，D（GPT-3.5）の自動評価と教師評価の相関係数は，いずれも「比較的強い相関がある」（0.4～0.7）ことを示している．D（GPT-4o）の自動評価と教師評価の相関係数は，「強い相関がある」（0.8 以上）ことを示している．

表 6　自動評価と教師評価の相関係数（N=129）

	Model C	Model D（GPT-3.5）	Model D（GPT-4o）	教師評価
Model C	1	.409**	.417**	.433**
Model D（GPT-3.5）	.409**	1	.769**	.684**
Model D（GPT-4o）	.417**	.769**	1	.807**
教師評価	.433**	.684**	.807**	1

$^{**}p < .01$

表 7 は，129 件の英文の Level 1 から Level 4 の自動評価と教師評価の件数である．

表 7　2024 年度 7 月レベル別英文件数

	Level 1	Level 2	Level 3	Level 4	計
Model C（機械学習）	1	1	34	93	129
Model D（GPT-3.5）	0	14	20	95	129
Model D（GPT-4o）	0	9	40	80	129
教師評価	0	10	51	68	129

表 7 の教師評価で Level 1 の英文が 0 件である理由としては，英文の中にわずかでも Detail 文に相当する内容が含まれている場合には，Level 2 と評価するという教員評価の決め事があることが影響している．例えば次のような語数がわずかな英文では，自動評価システム（Model C）は Level 1 と評価しているが，教員評価では Level 2 となる．

自動評価システム（Model C）が Level 1，教師評価が Level 2 の英文

I'd like to go to Korea. There are three reasons. First，I like K-POP. If I can go to Korea，I want to go to the place rerated K-POP. Second，I want to eat authentic delicious Korean food. Third，Korea is easy to go from Japan even if it's your first abroad. This is the reason the place I'd like to go to the most.

表 8 は自動評価と教師評価の差によるエッセイの数である．Model D（GPT-4o）は 129 件のうち，108 件で自動評価と教師評価が一致しており，また差が 1 の評価が 21 件，合わせて 129 件（100%）が差 1 以内となった（表 9）．

表 8　自動評価と教師評価の差による英文エッセイの件数

教師評価との差	Model C	Model D（GPT-3.5）	Model D（GPT-4o）
0	77	90	108
1	50	39	21
2	2	0	0
3	0	0	0
合計	129	129	129

表 9　教師評価との差が 1 以下のエッセイ件数の割合

Model C	Model D（GPT-3.5）	Model D（GPT-4o）
98.4%	100%	100%

Model D では，エッセイ入力ボックスに英文を入力したときに「評価できませんでした」という表示が出るケースが数件発生した．しかし，時間をおいて再度試すと評価が可能となった[6]．

5. 2. 自動評価と教師評価の差が大きかったもの

Model C の自動評価と教師評価で評価 Level に 2 以上の差のあった英文は 2 件であった．Model D（GPT-3.5），Model D（GPT-4o）の自動評価と教師評価で評価 Level に 2 以上の差のあった英文は 0 件であった（表 10）．

表 10　評価 Level の差が 2 以上だった英文の自動評価と教師評価およびその差

教師評価―自動評価	Model C	Model D（GPT-3.5）	Model D（GPT-4o）
0	77	90	108
1	11	8	4
2	0	0	0
3	0	0	0
-1	39	31	17
-2	2	0	0
-3	0	0	0
合計	129	129	129

自動評価が教師評価より高いものは Model C で 41 件（1 点差 39 件，2 点差 2 件），Model D（GPT-3.5）で 31 件，Model D（GPT-4o）で 17 件であった．自動評価が教師評価より低いものは Model C で 11 件，Model D（GPT-3.5）で 8 件，Model D（GPT-4o）で 4 件であった．この結果から自動評価（特に機械学習による Model C）が教師評価より高い点数を出す傾向が読み取れる．

表 11 は差が 2 以上だった 2 件の英文の Model C の自動評価と教師評価である．どちらも自動評価が 4，教師評価が 2 であった．Model D では教師評価との差が 2 以上の英文はなかった．

表 11　評価 Level の差が 2 以上だった英文の自動評価 Model C と教師評価

No.	モデル	教師評価	自動評価	差
1	Model C	2	4	2
2	Model C	2	4	2

以下の 2 つの英文は自動評価（機械学習による Model C）が教師評価より 2 点高かった（自動評価 4，教師評価 2）例である．

【エッセイ 1】（教師評価：2, Model C：4, Model D（GPT-4o）：2）

以下のエッセイ 1 は，Model C は Level 4 であるが，理由に続く Detail 文の内容が少なく平凡であること，さらに，複数の文に動詞がない点や文頭が小文字となっている点などエッセイとし

ての欠点が目立つため教師評価を Level 2 としたケースである．Model C では文法エラーと機械的エラー（mechanics）が評価に反映されない例と考えられる．Model D（GPT-4o）では教師評価と一致しており，ChatGPT 使用の評価システムの評価精度が上回っていることが確認できるケースである．

The country I want to go to is Hawaii. first reason that I wanted to go to Hapuna Beach, which has been voted the most beautiful beach in America. second reason I love the food in Hawaii. I want to try the acai bowl, loco moco, and poke bowl at this place. Third, I like the atmosphere of the cityscape because it has a slow atmosphere. For these reasons, I wanted to go to Hawaii.（75 words）

以下のエッセイ 2 は，3 つの理由それぞれに最低限の Detail 文があるため，教師評価を Level 2 とした．しかし，いくつかのスペルミスと文構造上に問題のある文が散見され全体の理解を損ねるため，Level 3 以上の評価とはならなかった．一方 Model C では，内容がそれほど深まっておらずスペリングや文法の誤りも多いにもかかわらず最高評価 Level 4 であるが，理由は不明である．Model D（GPT-4o）では教師評価との差は 1 点で，ChatGPT 使用のシステムの精度が上回っている．

【エッセイ 2】（教師評価：2, Model C：4, Model D（GPT-4o）：3）
The place I would like to visit the France. There are three reasos.

First, I want to eat the French Cuisine. I think good the taste because real French Cuisine is sure to be delicious.

Second, I want to visit the Palace and Park of Versaillse in Paris. I love rhe Manga “Lady Oscar”. I would love to see where there was used for that model.

Lastly, I want to go to Louvre Museum because I want to watch the “Mona Lisa” in Louvre Museum.

For there reasons, I would like to visit France the most.（96 words）

表 12 と表 13 は，機械学習モデル（Model C）と GPT-4o モデル（Model D）の教師評価との差が 1 のものの数を示している．表 12 は教師評価が高いケースで，表 13 は教師評価が低いケースである．

表 12　教師評価の方が Model 評価よりも 1 点分高いケース

Model C			Model D		
Model C	教師	教師評価の方が高い	Model D	教師	教師評価の方が高い
Level 3	Level 4	10 件	Level 3	Level 4	3 件
Level 2	Level 3	0 件	Level 2	Level 3	1 件
Level 1	Level 2	1 件	Level 1	Level 2	0 件
		11 件			4 件

表 13　教師評価の方が Model 評価よりも 1 点分低いケース

Model C			Model D		
Model C	教師	教師評価の方が低い	Model D	教師	教師評価の方が低い
Level 4	Level 3	33 件	Level 4	Level 3	15 件
Level 3	Level 2	6 件	Level 3	Level 2	2 件
Level 2	Level 1	0 件	Level 2	Level 1	0 件
		39 件			17 件

Model C，Model D ともに教師評価より高い評価が多いことから（39 件と 17 件），自動評価システムは人間による評価より高い評価を出す傾向がある．また，教師評価より低い場合でも高い場合でも，差のある評価は Level 3 と Level 4 の間で多く生じており，逆に低い評価（Level 1，Level 2）の間ではわずかとなっている．実際，教師評価時にも Level 3 と Level 4 の評価で判断に迷うことが多くあり，最高点とそれより 1 低い評価の判断の難しさが共通していることがわかる．

5. 4. 自動評価システムを利用した学生のアンケート分析

自動評価システム Model D（GPT-4o）を 2024 年度大学 1 年生に使用させ，アンケート調査を行った．学生には 2024 年 7 月の前期最終授業のライティングテストで作成した英文を，9 月の後期初回授業中，自動評価システムで評価させた．自動評価システム使用後に行った記述式アンケート結果についてテキストマイニング[7] を行い，自由回答の頻出語として抽出された単語同士の関係性を可視化するために共起ネットワーク分析を行った．テキストマイニングには，KH Coder 3 を使用した[8]．アンケートの質問は以下の 3 問である．

問 1．今回の自動評価システムについて良いと思う点を 2 つ書いてください．
問 2．今回の自動評価システムについて悪いと思う点を 2 つ書いてください．
問 3．今回のように自動評価システムを使って英語学習をすることが役に立ったか感想をお聞かせください．

5. 4. 1.　自動評価システムの良いと思う点

質問 1「今回の自動評価システムについて良いと思う点を 2 つ書いてください」に対する学生の自由記述回答の総抽出語数は 2,140 語（195 文）であった．抽出語の頻出語上位 5 件は「自分」（51 回），「評価」（43 回），「英文」（36 回），「サンプル」（25 回），「文章」（21 回）であった．

語の取捨選択を行わずに共起ネットワークを作成し，さらに共起性の強い線だけの描画に絞る「最小スパニングツリーだけを描画」を選択したところ，自動評価システムの良いと思う点に関して 8 つのサブグループが形成された（図 4）．

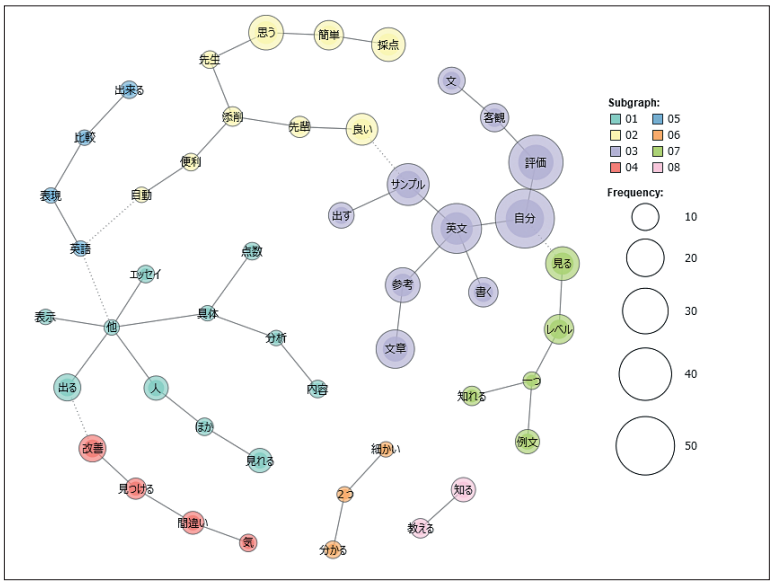

図 4　自動評価システムについて良いと思う点

一番大きいサブグループでは，「英文」，「自分」，「評価」，「客観」，「サンプル」，「参考」が示されていることから，学生は自分の英文の評価を客観的に見られること，またサンプル英文が参考になることを良い点としていることがわかる．次のサブグループには「評価」，「簡単」，「添削」，「便利」，「自動」，「先輩」，「良い」が示されており，自動評価が簡単で便利であり先輩のサンプルが出ることを良い点としている．以下はコメントの抜粋である．

「客観的に見た自分の文章の読みごたえを知ることができること．」
「自分の書いた英文の評価だけでなく，先輩の英文も見ることができるのでこういう文法の使い方もあるんだなと学ぶことができる」

3 つ目のサブグループでは「内容」，「分析」，「具体」，「点数」，「他」，「人」，「エッセイ」が示され，内容分析と具体的に点数が出ることを良い点としていることがわかる．4 つ目のサブグループには「見る」，「レベル」，「一つ」，「例文」，「知れる」が示され，一つ上のレベルの例文を

見ることを良い点としている．５つ目のサブグループには「改善」,「見つける」,「間違い」,「気(づく)」が示されていることから，自分では気づかない間違いを見つけ改善できることを良い点としていることが推測される．以下は上コメントの抜粋である．

「即座に点数などといった具体的な評価を提示してくれるのでいいと思いました.」
「自分より一つ上のレベルの例文を見ることができるので英文の改善につながると思った」
「自分では気づけない間違いを見つけることができる」

5.4.2. 自動評価システムの悪いと思う点

質問２「今回の自動評価システムについて悪いと思う点を２つ書いてください」に対する学生の自由記述回答の総抽出語数は 1071 語（159 文）であった．抽出語の頻出語上位５件（同率３位２件を含む）は「評価」(26 回),「改善」(9 回),「具体」(8 回),「評価」(8 回),「書く」(7 回）であった.

語の取捨選択を行わずに共起ネットワークを作成し，さらに共起性の強い線だけの描画に絞る「最小スパニングツリーだけを描画」を選択したところ，頻出語上位５件の語が現れなかった．線として描くべき共起関係がなかった語はネットワークに出ないため，この５件が現れるように「出現数による語の取捨選択」で「最小出現数」を４回に設定したところ，自動評価システムの悪い点に関して４のサブグループが形成された（図５).

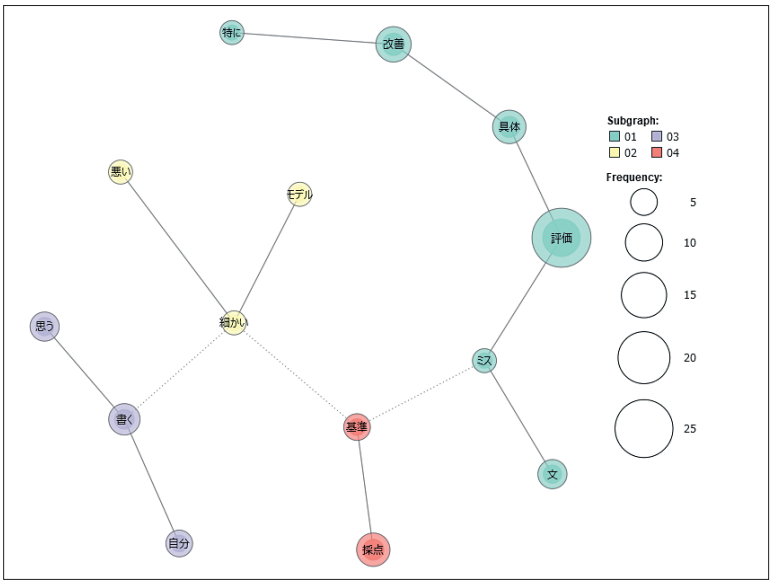

図５　自動評価システムについて悪いと思う点

1つ目のサブグループでは，「評価」，「具体」，「改善」，「ミス」，「文」が示されていることから，評価に具体性がなく，スペルミス，文法ミスが提示されないため改善できない点を悪い点と思っていることが推察される．2つ目のサブグループには「細かい」，「モデル」，「悪い」が示されており，もっと細かく評価してほしい，モデル D が評価してくれなかった，悪いところがわからないなどの点を指摘している．3つ目のサブグループでは「自分」，「書く」，「思う」が示されており，もっと細かく修正点を書いてほしいことが推察される．4つ目のサブグループには「評価」，「基準」が示され，自動評価システムの評価基準に疑問を持っていることが分かる．

「具体的にどんなところがだめでその評価なのか詳しく知りたい」
「4 月も 7 月も不安な文に仕上がったが，両方の評価が 4 であるため評価基準がわからない．話題が多くあれば文法的なミスがあってもよい評価を得られてしまうのかが気になる．」
「評価の際にどこが悪かったのか，どこがよかったのかについてのコメントがないため修正する箇所がわからない点．」
「モデル D で評価できなかった」[6)]

5. 4. 3.　自動評価システムを使って英語学習することが役に立ったかの感想

質問 3「今回のように自動評価システムを使って英語学習をすることが役に立ったか感想をお聞かせください」に対する学生の自由記述回答の総抽出語数は 2,859 語（158 文）であった．特徴を読み取りやすくするために「思う」「書く」という一般的な語を「語の取捨選択」で「使用しない語」に指定したところ，抽出語の頻出語上位 5 件は「自分」(61 回)，「評価」(47 回)，「英文」(38 回)，「役に立つ」(35 回)，「良い」(16 回）であった．

さらに共起性の強い線だけの描画に絞る「最小スパニングツリーだけを描画」を選択したところ，自動評価システムが役に立ったかの感想に関して 7 のサブグループが形成された（図 6）．

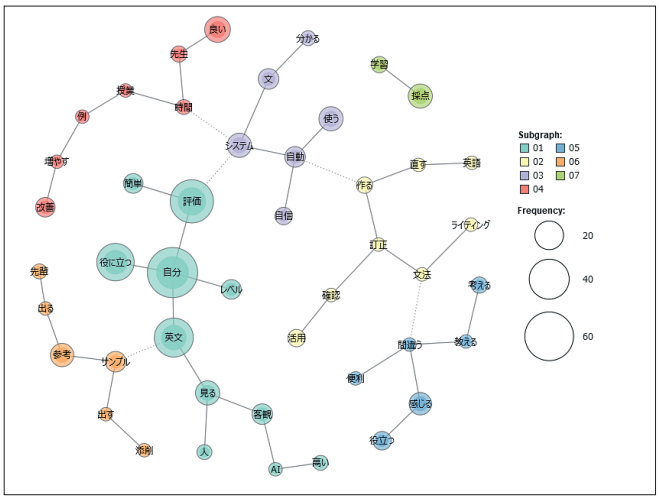

図 6　自動評価システムを使って英語学習することが役に立ったかの感想

一番多い1つ目のサブグループでは,「自分」,「評価」,「簡単」,「役に立つ」,「英文」,「見る」,「客観」,「AI」,「高い」が示されていることから,自分の英文がAIによって簡単に客観的に評価されることが役立ったと感じた学生が多かったことがわかる.次に多い2つ目のサブグループには「自動」,「システム」,「使う」,「自信」,「文」,「分かる」が示されており,この自動評価システムを使うことで自分の英文があっているかわかり自信がついたと感じていることが推察される.3つ目のサブグループでは「良い」,「先生」,「時間」,「授業」,「例」,「増やす」,「改善」が示され,授業で先生の時間を使うことなく表示される先輩の具体例をみて改善したり文の量を増やしたりできることが役だったことがわかる.以下はコメントの抜粋である.

「自分では客観的に判断することができないのでAIの技能をよりよく正しく使うことで,自分の能力をより高くしていければよいと思いました.」
「自分で英文を作って終わりだと不安も残るし,訂正すべき場所に気づけないまま終わることが多いので,自動評価システムを使うことで自信をつけられると思いました.また,最終的な英文のレベルを簡潔で的確に評価してもらえるのでわかりやすかったです.」
「たとえ授業中だったとしても先生が一人一人の文を読んで指導するのは難しく時間もかかるのでこのような自動評価システムの導入は授業の効率化につながると思う.」
「先生の手を煩わせることなくいつでも気軽に自分のライティングを客観的に評価してくれるのが良いと思いました」
「役に立ちました.先生に頼んで添削してもらおうとすると,お互い時間を使ってしまうので,このようなシステムがあればその場で評価できるから時短になってとても良いと思いました」

4つ目のサブグループには「英語」,「直す」,「作る」,「訂正」,「確認」,「活用」,「文法」,「ライティング」が示され,ライティングテストに向けてこのシステムを活用し文法力の改善にも役立ったと感じたことが推察される.5つ目のサブグループには「感じる」,「役立つ」,「間違う」,「便利」,「教える」,「考える」が示され,間違いがどのように間違っているかを考えることに役立ったことがわかる.6つ目のサブグループには「参考」,「サンプル」,「出す」,「添削」,「先輩」が示され,先輩のサンプル英文が参考になっていることが推察される.以下はコメントの抜粋である.

「自動評価システムを使うことによって,ライティングの基本や文法の基本を考えることができました.また,サンプルを表示してくれるので点数の取れる英文の書き方を知ることができました.」
「自分1人で作業しているときにでも活用できてとても便利だと思いました.間違ったところをどう間違っているのかと正解を提示してくれることがよかったです」
「機械的でも点が出るので話を広げたり,深めようとする意識が芽生えると感じた.先輩の

文章を読むこともよい刺激になった」

6. 考察

本研究では 2020 年度 1 回目から 2023 年度 2 回目までの計 11 回分のライティングテストの英文 1611 件を入力データとして生成 AI を利用した自動評価システム Model D を開発した．またそれを 2024 年度英語必修科目の授業内で大学 1 年生[5] に使用させた．この自動評価システムは，特定の教育現場で用いるための自動評価システムである．Model D 以前に使用した Model C は Amazon のクラウド上のサービスを利用し，一方 Model D は，生成 AI の GPT-4o を評価に使用したモデルである．どちらも技術者の協力により開発された自動評価システムである．

自動評価システムで評価する項目は英文の「内容の質」に限定している．すなわち主張を補足し説明する Detail 文の充実度によって Level 1 から Level 4 までの評価を行うシステムである．ライティング評価では欠かせない文法エラーやスペリングエラーは，この自動評価システムでは取り上げていない．学生たちには 2024 年度前期最終授業で実施したライティングテストの自分の英文を Model D に入力して評価結果を確認させた．

リサーチクエスチョン（1）「ChatGPT と機械学習による「内容の質」を測る自動評価システムはどの程度の精度があるか」に関しては，2024 年度の英文の計 129 件については，機械学習の Model C と教師評価の一致率は 59.7% であった．Model C の自動評価と教師評価の相関係数は，.433（比較的強い相関がある）で，1%水準で有意である．一方，生成 AI を使用した Model D と教師評価の一致率は，GPT-3.5 で 69.8%，GPT-4o で 83.7%であった．Model D の自動評価と教師評価の相関係数は，GPT-3.5 で .684（比較的強い相関がある）であり，GPT-4o では .807（強い相関がある）で，どちらも 1%水準で有意であった．

機械学習の Model C は 129 件のうち，77 件で評価結果が教師評価と一致しており，また差が 1 の評価が 50 件，合わせて 127 件（98.4%）が教師評価との差 1 以内に収まった．Model C の自動評価と教師評価で差が 2 以上の英文は 2 件のみであった．2 件どちらも Model C の評価が Level 4，教員評価が Level 2 であった．

生成 AI の Model D は 129 件のうち，GPT-3.5 では 90 件，GPT-4o では 108 件が教師評価と一致しており，また差が 1 の評価が，GPT-3.5 では 39 件，GPT-4o では 21 件であり，GPT-3.5 でも GPT-4o でも 100%が教師評価との差 1 以内に収まった．

よってリサーチクエスチョン（1）については，生成 AI を用いた評価システムの方が，機械学習を用いた評価システムよりも精度が高いことが明らかになった．今回の調査により GPT を用いた自動評価システムは一定の精度と信頼性を持つことが明らかになった．また，GPT-3.5 モデルより GPT-4o モデルの方が完全一致率が高く，生成 AI の大幅な性能向上に伴って精度が上がっていることがわかる．ただし，人間の評価との差が 1 以内の一致率は 100%ではあったものの，完全な一致は達成できておらず，これまでの研究者たちが指摘しているように，人間による評価と併用しフィードバックを補いながら活用することの重要性を再認識した．今後も目覚ましい生成 AI の発展に伴い，さらに精度の高い自動採点システムが開発されることが期待される．

リサーチクエスチョン（2）「自動評価システムを利用した学生のアンケートから示唆されるものは何か」については，3項目の記述アンケートから得られた学生のコメントについてテキストマイニングにより内容分析を行った．

まず，自動評価システム（Model D）について良いと思う点として，多くの学生が自分の書いた英文の評価が客観的・具体的・即時的に見られることと挙げている．同時に，評価とともに表示される先輩のサンプル英文から様々な気づきがあり，自分の英文の改善に役立ったと感じていることが伺える．気づきは，自分の間違い，新しい英語表現，文法の使い方など多岐にわたり，先輩のモデル英文が形成的評価ツールとして機能していることが確認できた．

一方，悪いと思う点としては，スペリングや文法の間違いが指摘されないことが挙げられた．この点に関しては，この自動評価システムでの評価は英文の「内容の質」に限定しており，文法エラーやスペリングエラーは，この自動評価システムでは評価しないという点を学生が理解していないことによるものであるため，使用前に「内容の質」に限定した評価点であることを周知しておく必要がある．また，自分の英文のどこが悪くどこが良いのかについて具体的な指摘がないという意見も見られ，先輩のサンプル英文から自分の英文を分析し改善点を見つけられない学生もいることがわかった．サンプル英文と自分の英文の比較分析できる力を養う指導も今後の課題としたい．また，自信のなかった英文が高評価だったため評価基準を疑問に思うという意見もあり，この点に関しては，多くの研究者が指摘しているように，自動評価システムは人間の評価者との完全な一致を達成するまでには至らないため，人間による評価と併用していく必要があることを改めて認識した．

最後に，自動評価システムを使って英語学習することが役に立ったかの感想については，ライティングテストのための準備をする過程で，不安に思っていた自分の英文を客観的に評価されることが改善と自信につながったという意見が多く見られた．中でも，先生に添削してもらわなくても自習時にその場ですぐに評価と模範例が出るため学習の効率化と時短につながる，というまさに自動評価システム導入のメリットを指摘する意見もあった．今後はこの自動評価システムをさらに効果的に授業に取り入れて，学生の英語力向上に役立たせたい．

今回開発した自動評価システムに関する学生のアンケート結果から多くの示唆を得ることができた．ライティング研究において「内容の質」を正確に評価することの難しさが指摘されてきた．本研究ではそれを乗り越えることを目的として，約3年にわたり自動評価システムの開発と実践を試みてきた．過去2年の機械学習による自動評価システムには精度に不足があったが，あらたに出現した生成 AI を使って開発を試みたところ教師評価との一致率は大きく向上し，学生のライティング力改善にも役立ったことがわかった．自動評価システムの開発は新たな時代に入り，今後のさらなる進歩を期待したい．

7. 終わりに

短期大学英語必修科目のライティングテストの英文データを用いて，生成 AI モデルによる自動評価システムを開発し，前年度まで使用していた人工知能の機械学習モデルによる自動評価シ

ステムとの精度の比較と，生成 AI モデルの信頼性の調査分析を行った．これまで英文ライティング指導でネックとなっていた教師の評価負担を軽減し，また学生が自ら英文を修正する動機付けとすることを目的として，2020 年度より収集した学生の英文エッセイ 1611 件をデータとして用いて開発した生成 AI モデル（Model D）を，それ以前の機械学習モデル（Model C）と比較した結果，教員評価との一致率も相関係数も生成 AI モデルが上回ることが確認された．また自動評価システムの使用が学生の英文ライティング力向上に活かされていることが示唆された．今後も自動評価システムが学生にとってより使いやすく有効なものとなるよう検討を続けていきたい．

謝辞

本研究は科学研究助成基金基盤研究©（課題番号 18K00814）の助成を受けたものである．本研究を進めるにあたり，株式会社ルーティングシステムズの大庭裕司氏から数多くの技術的サポートやアドバイスを受けた．ここに感謝の意を表する．

〔注〕

1. Integrated English は，短期大学の全学必修英語科目で Integrated English a（1 年前期），Integrated English b（1 年後期）を開講している．前後期とも週 2 コマの授業で，そのうち 1 コマは日本人教員，もう 1 コマは外国人教員が担当する．
2. 委託した業者はオンラインで英文添削サービスを提供する WEB システム開発会社で，11 年の実績がある．本研究における機械学習を使用した Model A, B, C および生成 AI を使用した Model D の開発は，同社に委託した．
3. Model A, B, C は「機械学習」の手法に沿って，学習用（70%）と検証用（30%）のデータを用意した．一方生成 AI モデルの教員評価との一致率は，GPT-3.5 の場合も GPT-4o の場合も，英文 100%（全 1611 件学習用）を学習したモデルの方が，英文 70%を学習したモデル（1128 件学習用，483 件検証用）より高い値を示していた（GPT-3.5 モデルの教師評価との一致率：100%学習モデルが 69.8%に対し 70%学習モデルは 62.0%，GPT-4o モデルの教師評価との一致率：100%学習モデルが 83.7%に対し 70%学習モデルは 67.4%）．そのため，生成 AI モデルによる調査に際しては，英文 100%を学習したモデルのみで調査を行うこととした．表 4 以降の結果はすべて英文エッセイを 100%学習した Model D によるものである．
4. CEFR とは，Common European Framework of Reference for Languages の略称である．CEFR A2 は，学生が入学時に受検する GTEC Academic の結果に基づいた英語レベルである．GTEC®（ジーテック / Global Test of English Communication）とは，株式会社ベネッセコーポレーションが実施している英語力を測定するためのスコア型英語 4 技能検定である．
5. 大学の組織改編により 2024 年度から短期大学生ではなく大学 1 年生がライティングテストを受けている．また Model D の使用についての感想も，大学 1 年生が行っている．大学 1 年生の英語力は，前年度の短期大学生と変わらず大部分が CEFR A2 レベルである．
6. Model D は ChatGPT からの 4 段階での回答を取得・表示させることを目的としているため，ChatGPT から 4 段階以外の別の評価結果が出力された際には，既存の評価数に誤評価が含まれてしまうことを防ぐため「評価できませんでした」としてエラーメッセージを表示させる仕組みになっている．
7. テキストマイニングは，文章データを単語ごとに切り取り，量的な方法で分析し，その結果を視覚化する内容分析の手法である．
8. KH Coder とは，計量テキスト分析またはテキストマイニングのためのフリーソフトウェアである．

〔参考文献〕

石井雄隆・近藤悠介．(2020)（編）『英語教育における自動評価 --- 現状と課題』．ひつじ書房．
岩田貴帆．(2020)．協議ワークを取り入れたピアレビューによる学生の自己評価力向上の効果検証．*大学教育学会誌，42*(1)，115–124.
小林雄一郎（2017）．「英語の自動作文評価」李在鎬（編）『文章を科学する』（pp. 158–174）．ひつじ書房．
丹原惇・斎藤有吾・松下佳代・小野和宏・秋葉陽介・西山秀昌．(2020)．論証モデルを用いたアカデミック・ライティングの授業デザインの有効性．*大学教育学会誌 = Journal of Japan Association for College and*

University Education, 42(1), 125–134.
三田薫・霜田敦子. (2023a). 英語初級学習者のパラグラフ・ライティングのための自動評価システム開発の試み. *実践女子大学短期大学部紀要 = Jissen Women's Junior College Review, 44*, 39–67.
三田薫・霜田敦子. (2023b). 学生の英文ライティング力向上の分析 その4：ルーブリックを用いた指導と「内容の質」を測る自動評価システム導入によるライティングの変化. *Jissen English communication, 53*, 2–35.
三田薫・霜田敦子. (2024). 英語初級学習者のパラグラフ・ライティングのための自動評価システム開発の試み Part 2. *実践女子大学短期大学部紀要 = Jissen Women's Junior College Review, 45*, 59–86.
Almusharraf, N., & Alotaibi, H. (2022). An error-analysis study from an EFL writing context: Human and Automated Essay Scoring Approaches. *Technology, Knowledge and Learning*, 1–17.
Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater® v.2.0. *ETS Research Report Series, 2004*(2), i–21. 10.1002/j.2333-8504.2004.tb01972.x.
Baffour, P., Saxberg, T., Abboud, R., Boser, U., & Crossley, S. (2024). Assessing ChatGPT's Writing Evaluation Skills Using Benchmark Data. *The Learning Agency*. https://the-learning-agency.com/insights/assessing-chatgpt-writing-evaluation-skills-using-benchmark-data/
EduKitchen. (2023, January 21). Chomsky on ChatGPT, education, Russia and the unvaccinated [Video]. https://www.youtube.com/watch?v=IgxzcOugvEI
Essel, H. (2023). 7 things you should know about ChatGPT. *BELI*. 10.17605/OSF.IO/AGWEQ .
Goodall Reece (2023) Australian universities to return to 'pen and paper' exams after student AI use. https://theboar.org/2023/01/australian-universities-to-return-to-pen-and-paper-exams-after-student-ai-use/
Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1–18.
Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *Relc Journal, 54*(2), 537–550.
Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050
Mizumoto, A., Shintani, A., Sasaki, M., & Teng, M. (2024). *Testing the Viability of ChatGPT as a Companion in L2 Writing Accuracy Assessment*. IRIS Database. https://doi.org/10.48316/q3zGE-HSpAC
Nature Editorial. (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature, 613* (7945) 612–612.
Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, Article 107769582211495. 10.1177/10776958221149577
Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, 2(3), 100083. https://doi.org/10.1016/j.rmal.2023.100083
Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford review of education, 13*(2), 191–209.
Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.
Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, 100270.
Uchida, S. (2024). Evaluating the Accuracy of ChatGPT in Assessing Writing and Speaking: A Verification Study Using ICNALE GRA. *Learner Corpus Studies in Asia and the World, 6*, 1–12. https://doi.org/10.24546/0100487710
Wiseman, C. S. (2012). A Comparison of the Performance of Analytic vs. Holistic Scoring Rubrics to Assess L2 Writing Iranian. *Journal of Language Testing, 2*(1), 59–92.
Zribi, R. & Smaoui, C. (2021). Automated versus Human Essay Scoring: A Comparative Study. *International Journal of Information Technology and Language Studies (IJITLS)*, 62–71.