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採点システム開発の試み

―英文の内容の質に特化した学習モデルの作成手法

An Attempt to Develop an Automated Scoring System for Paragraph Writing of Pre-intermediate Learners of English
―A Method for Creating a Learning Model Specific to the Quality of Content of English Writing

MITA Kaoru
三　田　　薫
英語コミュニケーション学科教授
SHIMODA Atsuko
霜　田　敦　子
英語コミュニケーション学科非常勤講師

抄録：

短期大学１年生向け英語必修科目で実施しているライティングテストのデータを用いた自動採点システムを開発した．人工知能の機械学習の「教師あり学習」に，過去の同一テーマのライティングテストの英文データを，教師評価と共に入力し，クラウドコンピューティングで一般公開されている人工知能の機械学習のサービスを用いて「教師あり学習」を行った．自動採点システムは英文の「内容の質」のみを採点対象とし，Level 1 から Level 4 の４段階の採点を行う．このシステムを用いて 2022 年度の英語必修科目のライティングテストの英文 124 件を採点したところ，自動採点システムと教師評価の一致率は 60.5%，相関係数は 0.544＊＊であった．同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 scoring system using data from past writing tests on one theme in a required English course for first-year junior college students and teacher evaluations that were input into a cloud-based artificial-intelligence “supervised” machine-

learning system that scored the “quality of content” of the texts at four levels (1–4). Then, 124 writing tests for the required English course in AY 2022 were scored; the agreement rate between the automated scoring system and the teacher’s evaluation was 60.5% (R = 0.544**). Students used the system in class and completed a questionnaire survey that was analyzed by text mining.

キーワード：自動採点システム，第 2 言語ライティング，パラグラフ・ライティング，相関分析，テキストマイニング，人工知能，機械学習，教師あり学習，一致率，内容の質

Keywords：Automated Scoring System, Second Language Writing, Paragraph Writing, Correlation Analysis, Text mining, Artificial Intelligence, Machine Learning, Supervised Learning, Agreement Rate, Quality of Content

1. はじめに

ICT や人工知能の発達は，日常生活だけではなく英語の 4 技能を測る診断テストにも大きな変化をもたらしている．これまで「読む」，「聞く」といった受容能力に比べて困難とされてきた「話す」，「書く」といった産出能力の評価にも，AI を用いた評価の導入が加速している．日本英語検定協会（2018）が，ライティングとスピーキングで通常採点に加え 2019 年度から自動採点を並行的に導入していく予定であることを発表したことは記憶に新しい．特にライティング（従来型の手書き）においては，2019 年度第 1 回から，英検 1 級から 3 級まですべてのテストで自動採点が導入されるという．同協会では英検における AI を用いた自動採点の主な特徴として，1）品質を保持したままでの 24 時間稼働の実現，2）人間による通常採点を補完する採点精度の向上，3）採点時間の短縮→採点期間短縮の実現，4）無回答や白紙答案仕分けによる採点者の負担軽減という 4 点を挙げている（日本英語検定協会，2018）．

海外の動向を見ると，日本よりはるかに早い段階からライティングにおける自動採点の実用化が始まっている．TOEIC ライティングテストでは，「E メール作成問題」と「意見を記述する問題」というエッセイタイプの問題を Educational Testing Service で開発された e-rater が採点している．この e-rater（ETS）のみならず，エッセイタイプの試験についてはすでに多くの実用的なシステムが存在しているという．アメリカでは 2012 年に企業や研究者による最適モデルのコンペティションで，意見文，説明文，叙述文を含む 8 つの論題について，e-rater を含む 9 つの自動採点システムで採点の性能が競い合われ，自動採点システムが人間の評価者に比べても信頼できるレベルにあるという報告が行われている（cf. 石岡, 2020）．e-rater 2.0 の精度については，専門家と e-rater の評価の一致度は 87% から 94% であったという報告もある（Burstein, Kukich, Wolff, Lu, Chodorow, Braden-Harder, & Harris, 1998; cf. 石井 & 近藤, 2020a）．

入学試験においては，大学進学率の上昇に伴って，大量の学習者の言語能力を効率的かつ客観的に測定する技術が強く求められている今，自動採点システムの導入が相次いでいる．アメリカ

では，TOEFL iBT のような英語検定試験のみならず，GMAT や MCAT などの大学院進学試験にもライティングの自動採点システムが導入されている．また韓国では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いる．日本英語検定協会に自動採点システムを提供した中国企業 iFlytek 社は，英語スピーキング試験向けの評価エンジンを開発している．それが中国の複数の都市における高校入試・大学入試で全面的に利用されており，年間利用者数が 300 万人であるという（小林，2017）．

自動採点システムは，検定試験や入学試験の採点だけではなく，教育現場でも注目されつつある．産出能力の測定には，採点者によって採点が異なる問題や人的・時間的コストなどの問題がある．そうした問題を軽減・解決するための手段として自動採点システムが導入されると，人的，時間的コストが大幅に減少し，システムが常に一定の点数を産出してくれる上に，一度構築してしまえば，人間が評価を行う場合に比べてかなり少ないコストで試験が行えるというメリットがある．しかし実際に自動採点導入を試みた複数の研究で，コンピュータの評価と人間の評価が必ずしも一致せず，むしろ教師も学習者も，自動採点システムによる評価およびフィードバックを信頼していないという研究がある（Li, Link, Ma, Yang, & Hedleheimer 2014; Li, Link, & Hedleheimer, 2015; cf. 石井 & 近藤, 2020b）．

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）．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採点モデルを開発するというささやかな試みを開始し，目下その精度を上げるための試行錯誤の最中である．今回は自動採点システム開発に至るまでの経緯と，自動採点システム試作版を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自動採点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 節で予備調査の概要，第 5 節で調査方法，第 6 節で調査結果を述べ，第 7 節で考察を行い，第 8 節でまとめる．

2. 先行研究

自動採点システムの評価と人間の評価者による評価の比較には，これまで一致率，相関係数，信頼性係数などの指標が用いられてきた（小林，2017）．英語の検定試験ですでに実用化されている自動採点システムについては，多くの研究論文で人間の評価と自動採点の一致率がかなり高いと報告されている（Burstein, Kukich, Wolff, Lu, Chodorow, Braden-Harder, & Harris, 1998; Yoon, Evanini, & Zechner, 2011; Zechner, Higgins, Xi, & Williamson, 2009; cf. 石井 & 近藤 , 2020b）．一方，教室内で自動採点システムを利用する場合，教師評価と自動採点システムの評価の一致率や相関係数は必ずしも高くない（Li, Link, & Hgelheimer, 2015; Wang & Brown, 2007）．これには教師の評価の観点は利用している自動採点システムの評価の観点と必ずしも一致していないことが影響しているとの指摘がある（石井 & 近藤，2020b）．

現在最も有名な英文自動採点システムは，ETS 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，談話といったライティングの様々な側面を計量的に評価している．さらにこのシステムは Criterion というウェブベースのフィードバック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（小林，2017）．しかし Criterion の教室内での利用に関して Li, Link, and Hegelheimer（2015）が調査したところ，自動採点と教師採点の相関係数が 2 種類の作文についてそれぞれ 0.42, 0.12 という結果となった．これは検定試験などで実用化されている自動採点システムの研究報告における指標と比べるとかなり低い値である．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）の自動採点システムによる採点と人間の評価者による採点を比較している．EFL 学習者 197 名のエッセイ採点について，人間の評価者と 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た．また Zribi and Smaoui（2021）はチュニジアの中レベルの英語力の大学生 15 名のエッセイについて，Paper Rater という自動採点システムの採点が人間の評価者の採点を大幅に上回っていたと報告している．

小田（2017）は，コンピュータを利用した英作文支援ツールとしては中国でシェア No.1 を誇る Pigai（ピーガイ）[2] を紹介している．Pigai は Criterion の中国版ともいえるもので，中国では大学を中心に 11 万人以上の教員と 1300 万人以上もの学生が利用している（2017 年時点）．このツールに日本人学生の書いた英文を入力したところ，人間の評価で高評価の英文の順番と Pigai の採点による高評価の順番は一致していたという．しかしパラグラフ・ライティングの基本である「最初に結論を述べる」「パラグラフの最初にはトピックセンテンスを置く」などの構成に関しては，Pigai は正しく採点できていないことを明らかにした．

日本国内の自動採点システムについては，中谷（2019）が，CEFR-J[3] 用に構築された英文エッセイの自動レベル判定システムを，手動採点の結果と比較している．複数のテストについての合格・不合格の一致人数は，調査対象者の英語レベルが上がるにつれて上昇し，3 段階で下から 27%，54%，68% と一致率が上がっている．しかし手作業による判定と自動判定の結果には，どのテストにおいても相関関係が見られなかった．このように日本の英語学習者向け自動採点システムは開発途上にあるものの，中谷はこうした研究の重要性を強調し，以下の理由を挙げている．すでに海外では授業用としてライティングサポートツールの Criterion や Write & Improve などが自動採点に活用されているが，これらは様々な母語を持つ被験者の英文ライティング採点データを用いているため，日本の英語学習者向けには必ずしも実用的ではない．中谷は日本の英語学習者特有のエラーを反映した自動採点システムの開発の必要性を訴えている．

従来自動採点の評価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．信頼性については，機械が人間よりも高い信頼性を持つという主張は現在広く認められている．すなわち機械による評価は，同一の英文に対して常に同じ採点結果を出すということである．それに対して人間による評価の信頼性については，ハロー効果（顕著な特徴に引きずられて他の特徴についての評価がゆがめられる），シークエンス効果（直前に読んだ英文が評価に影響する），中心化傾向（評価尺度の中心に評価が引きつけられる），長時間の作業による疲労の影響など，多くの問題があることが指摘されている（小林，2017，p. 161）．

妥当性に関しては，機械が人間のように正しく判定することはできないという批判が繰り返されている．一方，人間の評価も常に妥当であるとは限らないという指摘がある．Powers, Escoffery, and Duchnowsiki（2015）は人間の評価者の評価と Criterion の評価を比較し，訓練を受けた評価者の中で模範となる者，訓練を受けた評価者の中で平均的な者，訓練を受けていない者の順で Criterion との一致度が下がることを報告している．この結果について石井・近藤（2020b）は，「利用する自動採点システムについて十分な理解をせずに教室内に導入することは，作文の良し悪しを共有しない採点者を教室内に招き入れていることと同義である」（p. 125）と述べ，教師が自動採点システムを十分に理解した上で適切な利用方法を採用することの重要性を指摘している．

近藤・石井（2017）は自動採点の精度を上げる試みとして，課題を制限するという方法を紹介している．彼らは大学の英語授業のクラス分け試験および到達度試験に使用する発話採点システムの導入可能性を探った．その際，「談話完成タスク」を用いて学習者の発話を制限したところ，人間の評価者による点数とシステムが算出した点数の一致率は 74%であったという．この結果に基づき近藤・石井は，発話の自由度をある程度制限することで，クラス分け試験や到達度試験で十分に実用可能な自動採点が実現する可能性を示唆している．ただし，このシステムを実用化に向けて開発する際には，必要とするデータの収集と採点のコストが膨大になるという問題を同時に指摘し，「今後，少ないデータを用いて精度の高いシステムを構築することが可能になる研究が待たれる」（p. 37）とコメントしている．

学習者コーパスを有効に生かすことによって自動採点の精度を上げる研究も進んでいる．小林（2020）は，コーパス研究から導き出された「Biber の言語項目」「Hyland のメタ談話標識」「Coh-Metrix の指標」の各項目を特徴量として用いた自動採点システムを紹介している．その上で小林は，自動採点の採点対象となる学習者のパフォーマンスを正しく評価するため，異なる発達段階にある学習者の言語的特徴を定量的に記述する学習者コーパス研究から得られる知見を特徴量の選択に活かすことの重要性を説いている．

自動採点は，多くの場合，人工知能の「機械学習」と呼ばれる方法を用いて，学習者のパフォーマンス能力を自動で測定している．近年はこの機械学習の中でさらに進化した「深層学習」を自動採点に用いる研究が登場している．深層学習をエッセイの自動採点に適用した初期の研究に Alikaniotis, Yannakoudakis, and Rei（2016）がある．彼らは入力を単語列，出力をエッセイのスコアとする深層学習モデルを構築し，英語母語話者（middle school の学生）のエッセイを対象として，人間の採点結果とのスピアマンの順位相関係数で 0.91 という驚異的な結果を達成した（永田，2020）．

エッセイの観点ごとの得点（analytic score）を自動採点する研究もある．Ke, Inamdar, Lin, and Ng（2019）は，説得型エッセイ（persuasive essay）におけるエッセイの質の一つである主題に対する主張の強さ（thesis strength）を測るシステムの開発を試みた．Arguability 等 10 項目からなるルーブリックを作成して機械学習を行い，3 人の訓練された評価者のスコアとの相関関係を調べ，thesis strength はエッセイの全体評価に大きな影響を持つことを示している．

ここまで英語教育における自動採点の研究を見てきたが，本稿で紹介するのは，上記の取り組みとは異なるアプローチによる自動採点システム開発の試みである．すなわち自動採点のシステム開発の技能を持ち合わせていない現場教師が，クラウドコンピューティングで一般公開されている機械学習のシステムを用いてモデルを構築し，それを授業で活用するという手法を紹介するものである．あくまでも教室で英文ライティング指導をするという視点から，教師の採点負担軽減と学生の自主学習促進を目指して，短期大学 1 年次英語必修科目という限定された使用環境での自動採点システムの開発を試みた．

筆者らは，Amazon 社がクラウドコンピューティングを通じて一般にサービスを提供している Amazon Machine Learning による機械学習を用いて自動採点システムを開発する試みを行った．先行研究で取り上げたような特徴量の導入は，本調査では行っていない．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（ただし学生には同じテーマであっても毎回違う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採点の対象とせず，「内容の質」に限定して 4 段階の評価を行う，といった取り組みを行っている．この方法を用いて 2020 年度，2021 年度の授業内でライティングテストをそれぞれ年 3 回行い，そのデータに教師の採点結果を「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テーマを毎回同一にする理由は，データ数を増やす目的のみならず，学生に同じテーマについて繰り返し取り組ませることによってパラグラフ・ライティングの構造や内容の深め方の学びを定着させたいという意図がある．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，学生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）の意味合いが強い．

次節以降，自動採点システム開発のための予備調査および開発されたシステム（Model A）の詳細，また同システムを用いて 2022 年度 7 月のライティングテスト英文を採点した際の自動採点と教師採点の一致度，同システムを 2022 年度授業で学生に使用させた後の学生アンケートの結果を報告する．

3. リサーチクエスチョン

(1)　学生のライティングテスト英文の教師による採点と自動採点システムによる採点の一致度はどの程度であるか

(2)　自動採点システムを利用した学生のアンケート回答から示唆されるものは何か

4. 予備調査

本稿の調査を行う前に，以下の 2020 年度と 2021 年度の英文データを用いて予備調査を行った．

予備調査で使用したデータ：英文エッセイ 995 件の内訳

① 2020 年度：5 月と 7 月と 2021 年 1 月の英文エッセイ計 483 件

② 2021 年度：4 月と 7 月と 2022 年 1 月の英文エッセイ計 512 件

初めに 995 件について生データのみを用いた機械学習を行った．結果は Accuracy 73.367%，F1 値 0.688 であった．データ数の増加による影響を確認するために生データ 995 件をコピーし計 1990 件のデータセットを用いてモデルの作成を実施した．これによりデータセットには同じデータが 2 件ずつ存在する．結果は，Accuracy 90.955%，F1 値 0.838 であった．次に，データの件数と類似データの量の関係性，および人工的にデータ量を増やした際の影響調査を目的とし，以下の調査を実施した．生データ 995 件をおよそ半数の 497 件にし，その 497 件をコピーして倍にした総数 994 件をデータセットとしてモデルの作成を実施した．データセットには同じデータが 2 件ずつ存在する．結果は，Accuracy 93.97%，F1 値が 1 であった．まとめると以下の表 1 のようになる．

表 1　データの種類による Accuracy と F1 値

	（a）生データ	（b）生データ × 2	（c）（生データ /2）× 2
データ総数	995	1990	994
Accuracy	73.367%	90.955%	93.97%
F1 値	0.688	0.838	1

表 1 の結果より，同じ文章を用いてデータ総数を生データの倍にした（b）では（a）に比べて Accuracy および F1 値が上がること，また生データのデータ件数を半分にしてそれを 2 倍した（c）が Accuracy 及び F1 値で最も高い値となることが明らかになった．（b）と（c）の比較によりデータの件数自体が少ないことでデータセットに含まれるエッセイの表現のパターンも減少し，それに伴って，適合率や再現率，また正解率が上昇し結果として F1 値や Accuracy が上昇したと考えられる．このことから，学習させるデータセット内のエッセイに含まれるエッセイのパターンの量すなわちデータセットに同じような文章が多く含まれるほど，Accuracy や F1 値が上昇しやすいことが予想される．

機械学習にかけるデータとしてエッセイの内容や形式に制限を設けないエッセイを用いる場合は，エッセイの表現パターンが大幅に増え，それに伴い機械学習に必要なデータ量が膨大になることが予想される．それを回避するため，入力データの条件を絞り，表現を一定のパターンに限定することにより，データ量を抑えながら，特定条件下でのみであるが，有効なモデルが作成できる可能性があると考えた．

予備調査ではまた，データ数を増やす別の試みとして，ひな形を数個用意して特定箇所（地名や固有名詞）を置換しデータ数を増やすという実験も行った．しかしこの場合，特定箇所のみ異なる文章が増え，データに偏りが生まれる．また，本来の手段で学生が作成した英文エッセイを生データとして使用する場合とは，データにズレが生じる．このような問題を避けるため，本調査では人工的に特定箇所を置換したデータではなく，学生が作成した生データを機械学習に用いるという方針で作業を進めることとなった．

5. 調査方法

5. 1. 自動採点システムの概要

今回開発した自動採点システムを Model A とする．Model A の概要，システム構築の前提条件，外部インターフェイスは以下の通りである．データ入力や外部インターフェイスのシステム構築は専門の業者[4] に委託した．

(1) 本稿で用いる自動採点システム（Model A）の概要

① AI の機械学習の「教師あり学習」を用いる．

② 各英文についての教師の評価を「教師あり学習」の「正解」として与える．

③ 「内容の質」について Level 1 から Level 4 を予測する「分類」を行う．

(2) 英文採点インターフェイスのシステム構築の前提条件

① API（Application Programming Interface）を使用して英文の採点結果を取得，表示する．

② 複数の評価モデルを切り替えて使用することを可能とする．

(3) 外部インターフェイス（WebAPI）

① Amazon Machine Learning エンドポイント

② テキストの送信と評価の取得に使用

(4) 機械学習に用いるデータ

2021 年度に 3 回実施したライティングテストの英文データと教師の採点データ

5. 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 2021 年度ライティングテストは，3 回とも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ブレインストーミングを含めて 15 分間であり，テスト用画面には，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．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[5]．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 want to visit most is（　　）. There are three reasons.

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．

5. 2. 学生英文の評価基準

機械学習の「教師あり学習」で用いる教師評価は，「内容の質」に関する以下の５つの評価基準のうち Level 1 から Level 4 の基準で評価した．「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 0：未完成，理由が３つないもの，順番のディスコースマーカーがないもの，単語羅列で意味不明のもの
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

「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，自動採点システムでは採点の対象としていない．以下は０から４の各レベルの英文サンプルである．

① 未完成のもの「Level 0」例
The place I want to visit most is Hawaii. There are three reasons. First, I want to swim because sea is very beautiful in Hawaii. Second, Hawaiian food is very yummy. For example, rokomoko, garlic shrimp.

② Detail 文が無いもの「Level 1」例

The place I want to visit most is China There are three reasons.
First, I want to feel the atmosphere. Second, I want to try speak Chinese. Third, I want to know Chinese culture.（34 words）

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 Kyoto. There are three reasons. The first reason is I love Kyoto. I enjoyed in Kyoto when I went to there two years ago. The second reason is I eat delisious foods, for example, yatsuhashi and matcha. I like matcha. The third reason is I am interested in Kyoto history. I want to see beautiful temples. For those reasons, I want to visit Kyoto.（72words）

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 want to visit most is Toyama. There are three reasons. First, I like the most is the scenic beauty. The air is clear, and I feel like it. I'm feeling better even if I'm worried or depressed. This will my point of view. Secondly, the rich nature and clean water make the food delicious. I've heard that the calf fish is a popular hidden fish. In addition the fish called "Fukurahagi" in japan. It is delicious even if you actually eat it. The third is where people are good. Many people are kind and cheerful. It makes me want to visit again.（105words）

④ 「3」の中で特に優れているもの「Level 4」例

The place I want to go most is Kamakura. There are three reasons.
The first reason is that I have been to Enoshima many times, but I have spent a day in Kamakura only once. Therefore, I have never walked around Kamakura from end to end. The last time I visited, I went to famous sightseeing spots such as Hase-dera, Komachi-dori, and Tsurugaoka Hachimangu, but I still haven't been to Komyoji and Myohon-ji. The second reason is that I want to go to a place I wanted to go to the last time I visited, but I couldn't. Actually, the last time I visited Kamakura with my friends, the Kamakura Museum of Literature was closed. I and my friends love literature very much, so I definitely want to go to the Kamakura Museum of Literature next time. The third reason is that I want to see the scenery of Kamakura at night. When I go to Enoshima or Kamakura, I often get tired at night and go home early. Some temples light up at night, so I'm very curious about how Kamakura's Japanese atmosphere and rich nature look like. That's why I want to go to Kamakura.（197words）

5. 3. 入力データとなる学生英文の見直し

予備調査の結果に基づき，機械学習に用いるデータは人工的・機械的に作成した文ではなく生データのみとする方針となった．それを受けて機械学習の「教師あり学習」で付与される評価レベルを見直すこととなった．

2020 年度以降のライティングテストでは，元々4 段階の評価レベルが設定され，一番下の Level 1 の基準は理由のみが書かれ「Detail 文が無いもの」となっていた．しかし Level 1 より下のレベルが設定されていなかったため，「Detail 文が無いもの」以外に以下の①から⑤のような雑多な要因により Level 1 の教師評価となっているものが多数あり，それが自動採点の精度を下げる可能性があると判断された．そこで未完成のもの，「第 1 に」「第 2 に」といった順番のディスコースマーカーが入っていないもの，3 つの理由を書くべきエッセイで 3 つ書かれていないもの，内容が理解できないものを Level 0 として，機械学習に用いるデータから外すこととした．

① 文の途中で終わっている（未完成）
② 順番のディスコースマーカーがない
③ 理由が 2 つ
④ 理由が 1 つ
⑤ 意味不明

この基準で 2021 年度データを改めて点検した結果，以下の表 2 のように計 512 件のエッセイのうち Level 0 のエッセイが 94 件（約 18.4%）も含まれていることが明らかになった．

表 2　2021 年度エッセイのうちのゼロ評価エッセイ件数の割合

	エッセイ件数	Level 0 エッセイ件数	割合 %
2021 年 4 月	171	47	27.5
2021 年 7 月	171	28	16.4
2022 年 1 月	170	19	11.2
総数	512	94	18.4

生データが元々少ない中，これら 94 件のデータが除外されると，採点精度を大きく下げる可能性がある．そこでエッセイの生データを減らさないための方策として，Level 0 のエッセイを Level 1 以上のエッセイにリライトする作業を英文添削の専門家に委託した[6)]．こうしたリライトエッセイは特定箇所（地名や固有名詞）を人工的に置き換えたデータと異なり，学生英文の生データと同等のものと筆者らは判断し，リライトエッセイ 94 件を含めた 512 件を機械学習の学習データとした．

5. 4.　自動採点システムの表示画面

今回開発した自動採点システム（Model A）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・送信後の画面である．

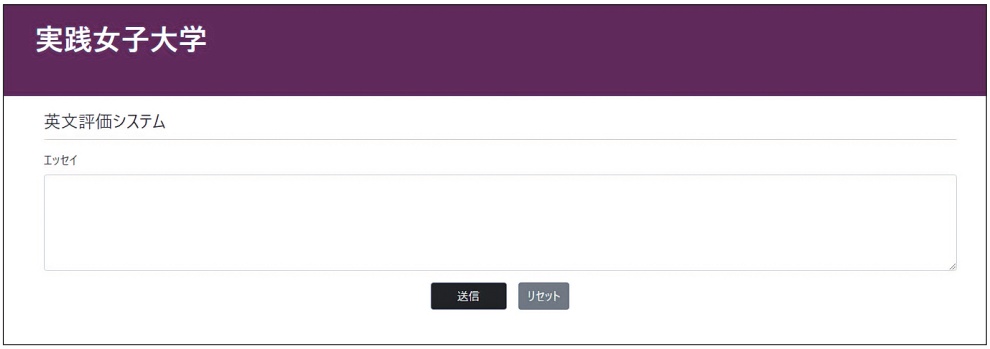

図 1　英文評価システムの英文入力前画面

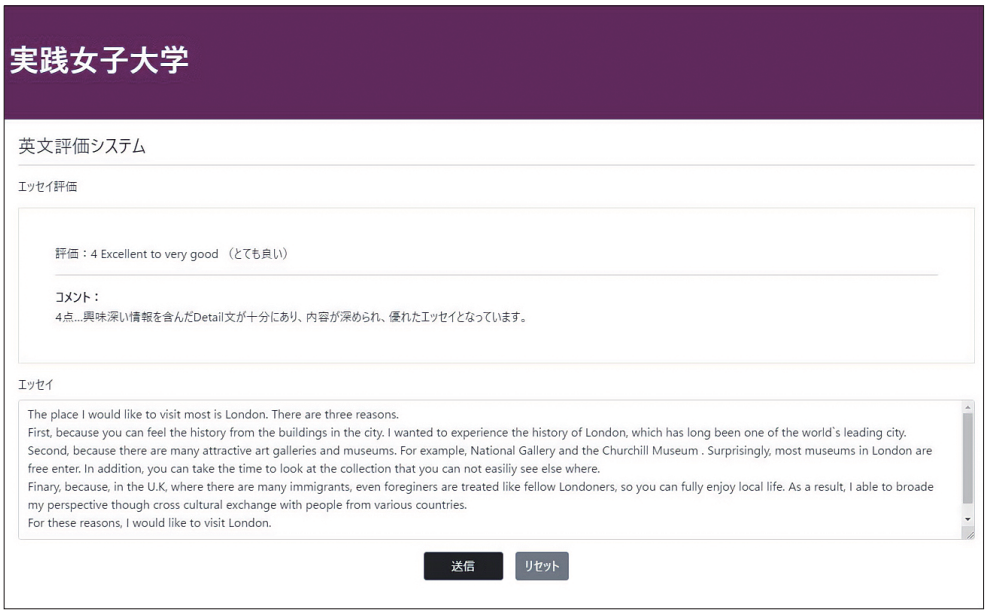

図 2　英文評価システムの英文入力・送信後画面

以下の表 3 は，英文入力・送信後に画面表示されるレベル別のフィードバック表現である．

表 3　自動採点システムに表示されるフィードバック表現

Level	自動採点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．

5. 5. 分析方法

2021 年度ライティングテストのデータを基に開発された自動採点システム Model A を，2022 年 9 月の授業で学生に使用させた．その上で以下の 3 つの分析を行った．

(1)　今回開発した Model A の自動採点と教師採点との一致率
(2)　自動採点と教師採点と差が大きかったものの理由を分析
(3)　学生アンケートの自由記述のテキストマイニング分析

6. 調査結果

6. 1. 自動採点と教師採点の一致率と相関係数

2022 年 7 月に実施したライティングテストのエッセイについて，9 月の後期授業開始時に Model A を使用させた．学生の英語レベルは CEFR A2 レベル中心である[7]．表 4 は，2022 年度学生英文 124 件の Model A の Level 1 から Level 4 の自動採点と教師採点の一致率である．

表 4　自動採点と教師採点の一致率

一致英文件数	英文総件数	一致率（%）
75	124	60.5

表 5 は，4 段階評価の自動採点と教師採点の相関関係，平均，標準偏差である．相関係数は 1%水準で有意である．4 段階評価の自動採点と教師採点の相関係数は，「比較的強い相関がある」（0.4～0.7）ことを示している．

表 5　自動採点と教師採点の相関係数と平均, 標準偏差（N=124）

	自動採点	教師採点	*M*	*SD*
自動採点	—	.544**	3.202	.9105
教師採点	.544**	—	3.056	.8954

**$p < .01$

表 6 は，124 件の英文の Level 1 から Level 4 の自動採点と教師採点の件数である．

表 6　各 Level の自動採点と教師採点の件数

評価 Level	自動採点の件数	教師採点の件数
1	4	5
2	29	31
3	29	40
4	62	48
合計	124	124

表 7 は自動採点と教師採点の差によるエッセイの数である．124 件のうち，75 件で採点結果が一致しており，また差が 1 の採点が 39 件，合わせて 114 件（92.0%）が差 1 以内となる．

表 7　自動採点と教師採点の差による英文エッセイの件数

採点の差	英文エッセイの件数
0	75
1	39
2	7
3	3
合計	124

6. 2.　自動採点と教師採点の差が大きかったもの

自動採点と教師採点で評価 Level に 2 以上の差のあった英文は 10 件であった．表 8 は 10 件の英文の自動採点と教師採点，およびそれらの差である．

表 8　評価 Level の差が 2 以上だった英文の自動採点と教師採点およびその差

No.	自動採点	教師採点	差
1	4	1	3
2	2	4	-2
3	4	1	3
4	4	2	2
5	4	2	2
6	4	1	3
7	1	3	-2
8	4	2	2
9	4	2	2
10	4	2	2

10 件の自動採点と教師採点の差が 2 以上のものの中で，3 点の差があったものが 3 件（自動採点 4，教師採点 1），2 点の差があったものは 7 件であった．また，差が 2 点だった 7 件のうち教師採点の方が自動採点より良かったものは 2 件，残りの 5 件は自動採点の方が教師採点より良かった．この結果から自動採点が教師採点より高い点数を出す傾向がわかった．

以下の英文は，自動採点と教師採点の差が最も大きかった（自動採点 4，教師採点 1）英文の例である．教師採点の内容評価の基準である「Detail 文」がなく，自動採点が「4」になった理由は不明である．

【自動採点が 4，教師採点が 1 の学生エッセイ】

I would like to go Hokkaido in Japan.I has to three reasons.
First, the fish and shellfish are very tasty.
Secondly, it is cooler in sumer.
Finally, I have never been to Hokkaido.
Threfore, I want to go Hokkaido.（39 words）

以下の英文は自動採点が教師採点より 2 点低かった（自動採点 2，教師採点 4）英文である．

【自動採点が 2，教師採点が 4 の学生エッセイ】

The place I would like to visit most is America. There are three reasons.
First, America has many delicious foods. In particular, America is famous is firstfood. I like firstfood. I want to eat a big hamburger. On the other hand, my mother doesn't like firstfood.

Second, There are many places I want to visit in America. For example, I want to visit Disney Land and Universal Studios. There are two Disney Lands in America. Disney Land in America is larger than Disney Land in Japan. I like to ride on attractions. I want to ride many attractions.
Finally, I want to go to Hollywood because I like watching Hollywood movies. If I go there, I want to see where they were filming.
For these reasons, I would like to visit America.（132 words）

この英文は 3 つの理由それぞれに複数の Detail 文があり，理由の具体例や詳細が記述されている．教師の内容評価の基準では明らかに「4」と判断されるケースである．

以下の英文は自動採点が教師採点より 2 点高かった（自動採点 4，教師採点 2）英文の例である．

【自動採点が 4，教師採点が 2 の学生エッセイ】
The place I would like to visit most is Australia. There are three reasons.
First, I love koaras and would like see Australian koaras in person.
Secound, I would like to visit Opera House because the night view of the Opera House is very beautiful.
Finally, I have many things to do in Ausuralia, for example, shopping and running.
For these reasons, I would like to visit Australia.（68 words）

この英文は，3 つの理由のうち一つ目に Detail 文がない．教師採点では理由 3 つにそれぞれ Detail 文がない場合「2」の評価となる．この英文では 1 つめの理由に詳細がないため「3」以上の評価が与えられなかった．このケースで自動採点が「4」になった理由も不明である．

6. 3. 自動採点システムを利用した学生のアンケート分析

自動採点システム（Model A）を 2022 年度学生に使用させる機会を設けた．同学生は 2022 年 7 月の前期最終授業でライティングテストを受けている．そのテストで作成した自分の英文を，9 月の後期初回授業中，自動採点システムで採点させた．自動採点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[8] 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[9]．アンケートの質問は以下の 5 問である．問 1 から問 4 は 7 月に学生が作成した自分の英文を自動評価システムにかけた後の質問で，問 5 は評価をさらに上げるように英文を修正させた後に再度自動採点システムを試させた後の質問である．

問 1. 今回の自動採点システムについて良いと思う点を 2 つ書いてください．良い点がなければ「ない」と書いてください．

問 2. 今回の自動採点システムについて悪いと思う点を 2 つ書いてください．悪い点がなければ「ない」と書いてください．

問 3. 今後，このシステムのどのようなところを改良すれば使ってみたいと思いますか．

問 4. 今回の自動採点システムについての感想をお聞かせください．

問 5. 修正英文についての自動採点システムの評価について，感想を自由に書いてください．

6. 3. 1. 自動採点システムの良いと思う点

質問 1「今回の自動採点システムについてよいと思う点を 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 2,200 語（193 文）であった．抽出語の頻出語上位 5 件は「評価」(37 回)，「採点」(35 回)，「自分」(35 回)，「結果」(19 回)，「書く」(18 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，自動採点システムの良い点に関して 6 つのサブグループが形成された（図 3)．

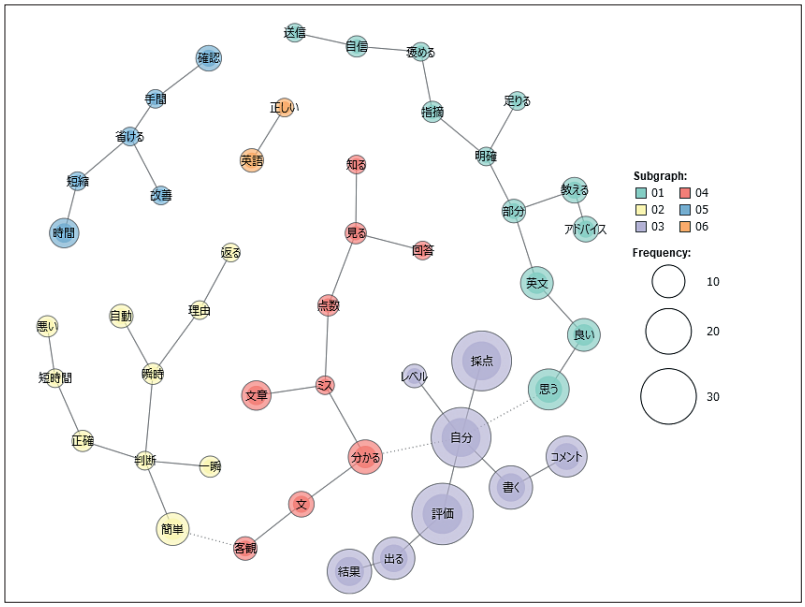

図 3　自動採点システムについて良いと思う点

1 つ目のサブグループでは，「英文」，「良い」，「部分」，「明確」，「指摘」が示されていることから，学生は英文の間違い部分が明確に指摘されていることを良い点としていることがわかる．2 つ目のサブグループには「簡単」，「判断」，「瞬時」，「自動」，「正確」が示されており，簡単に瞬時に自動的に正確に判断される点を良い点としている．3 つ目のサブグループでは「自分」，「採点」，「評価」，「出る」，「結果」，「書く」，「コメント」が示され，学生は自分の書いた英文にコメント付きで採点や評価が出ることを良い点としている．4 つ目のサブグループには「文章」，「ミス」，「分かる」，「文」，「客観」，「点数」，「知る」が示され，文章のミスが分かり客観的に点

数を知ることを良い点としている．5 つ目のサブグループには「時間」,「短縮」,「手間」,「省ける」,「改善」が示されていることから，時間が短縮され手間が省け改善されることを良い点としていることが推測される．以下は上位 5 サブグループのコメントの抜粋である．

1. 分析をおこなってくれるため何が間違ってしまったのか明確になる．英語に対するモチベーションになる．
2. 事前に書いておいた英文を張り付けるだけで，簡単に自分が書いた英文の質を見極めてくれるので，そこが良いと思います．
3. すぐに採点されるから効率的．評価だけでなくコメントもついているのがいい．
4. 自分の文章が，単語のミスや文法の間違いだけをチェックされるのではなくて，客観的に文を読まれた時の意見も聞けるのでとてもいいと思う．簡単に自分の文章の点数が分かるのが便利なのでいいと思った．
5. すぐに採点結果が出るので手間が省ける点．

6. 3. 2. 自動採点システムの悪いと思う点

質問 2「今回の自動採点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 1,160 語（273 文）であった．特徴を読み取りやすくするために「思う」という一般的な語を「品詞による語の取捨選択」で「使用しない語」に指定したところ，抽出語の頻出語上位 8 件（同率 4 位まで）は「評価」(15 回),「採点」(9 回),「ミス」(7 回),「アドバイス」(6 回),「悪い」(6 回),「改善」(6 回),「指摘」(6 回),「良い」(6 回）であった．

「思う」を外した共起ネットワークを作成し，さらに「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採点システムの悪い点に関して 7 つのサブグループが形成された（図 4）．

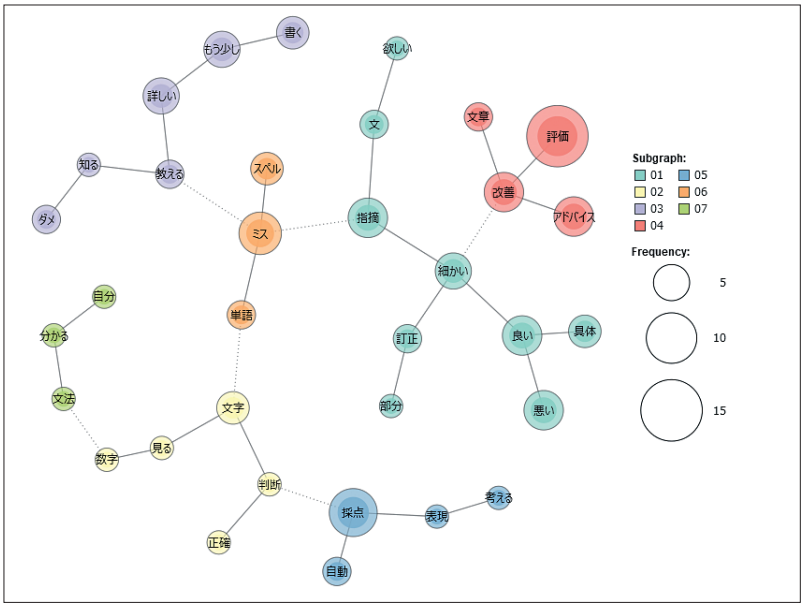

図４　自動採点システムについて悪いと思う点

１つ目のサブグループでは，「細かい」，「指摘」，「文」，「欲しい」，「訂正」，「部分」，「良い」，「悪い」，「具体」が示されていることから，悪い点として細かい指摘や訂正が欲しいことが推察される．２つ目のサブグループには「文字」，「判断」，「正確」，「見る」，「数字」が示されており，採点結果画面のフィードバック部分の文字が読みにくいという欠点を指摘している．３つ目のサブグループでは「詳しい」，「もう少し」，「書く」，「教える」，「知る」，「ダメ」が示されており，詳しいことをもう少し教えてダメな点を知りたいと思っていることが推察される．４つ目のサブグループには「評価」，「改善」，「アドバイス」，「文章」が示され，採点結果が改善されるようにするための具体的なアドバイスを求めていることが分かる．５つ目のサブグループには「採点」，「自動」，「表現」，「考える」が示され，コンピュータの自動採点では人間の採点に比べて見落とす点がある可能性があることを指摘している．６つ目のサブグループには「ミス」，「スペル」，「単語」が示され，これらは「指摘」，「教える」とも関連していることから，単語のスペルミスを指摘して教えてもらえないことを悪い点と考えていることが推察される．７つ目のサブグループには「文法」，「分かる」，「自分」が示され，文法を間違えた個所がわからないことを悪い点と考えていることが分かる．以下は各サブグループのコメントの抜粋である．

1. どこを間違えたかの細かい訂正，指摘がされないので結局のところ具体的改善方法がわからず，改善しづらい点．
2. 文字が全て同じ大きさのため評価の数字が見にくい点
3. どこがダメなのかより詳しく教えてほしい．
4. 採点して数字での評価がもらえてもその文章を改善するためのアドバイスやヒントがない

ため，低い評価をもらった時に改善できるのかと思ったこと．

5. 型にはまった採点になるために，見落とすミスが出てくることがあると考える．逆に，表現としては伝わるし間違っているわけではないのに間違いとされることがあると考える．
6. 単語のスペルミスがあったのに評価が 4 だったので，ほんとかなと思うところがあった．
7. 文法や単語を何処でミスしたかわからない点

6.3.3. 自動採点システムの改良点

質問 3「今後，このシステムのどのようなところを改良すれば使ってみたいと思いますか」に対する学生の自由記述回答の総抽出語数は 1,804 語（131 文）であった．特徴を読み取りやすくするために「思う」という一般的な語を「品詞による語の取捨選択」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「使う」(24 回)，「改善」(22 回)，「評価」(16 回)，「アドバイス」(15 回)，「コメント」(13 回）であった．

「思う」を外した共起ネットワークを作成し，さらに「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採点システムの改良点に関して 7 つのサブグループが形成された（図 5)．

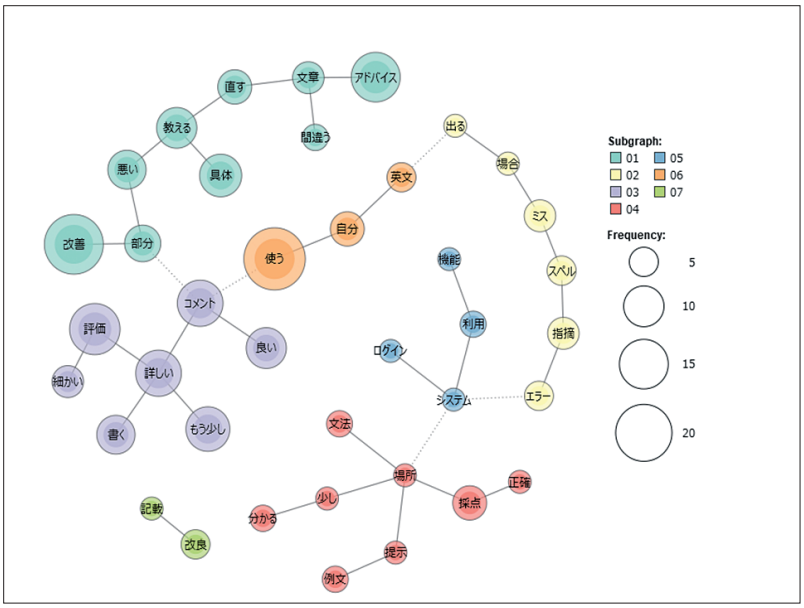

図 5　自動採点システムについて改良点

1 つ目のサブグループでは，「改善」，「部分」，「悪い」，「教える」，「具体」，「直す」，「文章」，「間違う」，「アドバイス」が示されていることから，改善すべき部分の悪い点を具体的に教え間違った文章を直すアドバイスを求めていることが分かる．2 つ目のサブグループには「エラー」，「指摘」，「スペル」，「ミス」が示されており，スペルエラーの指摘を望んでいることが推察される．3 つ目のサブグループでは「評価」，「細かい」，「詳しい」，「もう少し」，「コメント」，「良い」が示され，評価にもう少し詳しいコメントが書かれていると良いと思っていることが分か

る．4つ目のサブグループには「採点」,「正確」,「場所」,「提示」,「例文」,「文法」,「少し」,「分かる」が示され，より正確な採点や正しい例文の提示，文法の誤り箇所が分かるようになることを望んでいることがわかる．5つ目のサブグループには「利用」,「機能」,「システム」,「ログイン」が示され，利用する機能のシステムログインに改善を求めていることが分かる．6つ目のサブグループには「使う」,「自分」,「英文」が示され，自分の英文以外の良い例の紹介を期待していることが分かる．7つ目のサブグループには「改良」,「記載」が示され，改良点が記載されることを求めていることが推察される．以下は各サブグループのコメントの抜粋である．

1. 文章のアドバイスやヒントなどのコメントや，文法が間違っていたときにどこが間違えているのかがわかるようになれば沢山利用したい気持ちになる．
2. スペルミスや，字下げ，Because エラー等のミスをした場合，赤いライン等で指摘が入ると改善しやすいと考える．
3. もっと詳しく長文で評価のコメントを書くことを改善すれば使ってみたい．
4. 改善したほうがいいところやアドバイスがあればそれを参考にすることができると思うため，例文を提示してもらえるといいのではないかと感じた．
5. ログインの手間が面倒なので，それを無くして欲しい．そうすればもっとこのシステムを利用すると思う．
6. 自分が使った英文以外にもさらに良い表現を例として紹介してくださるともっと学べると思いました．
7. 何がよくて何が悪かったかということを明確に記載するように改良すれば使ってみたい．

6.3.4. 自動採点システムを使った感想

質問4「今回の自動採点システムについての感想」に対する学生の自由記述回答の総抽出語数は2,625語（147文）であった．特徴を読み取りやすくするために「思う」という一般的な語を「品詞による語の取捨選択」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「採点」(44回),「評価」(34回),「自分」(32回),「英文」(18回),「使う」(18回）であった．

「思う」を外した共起ネットワークを作成し，さらに「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採点システムを使った感想に関して5つのサブグループが形成された（図6).

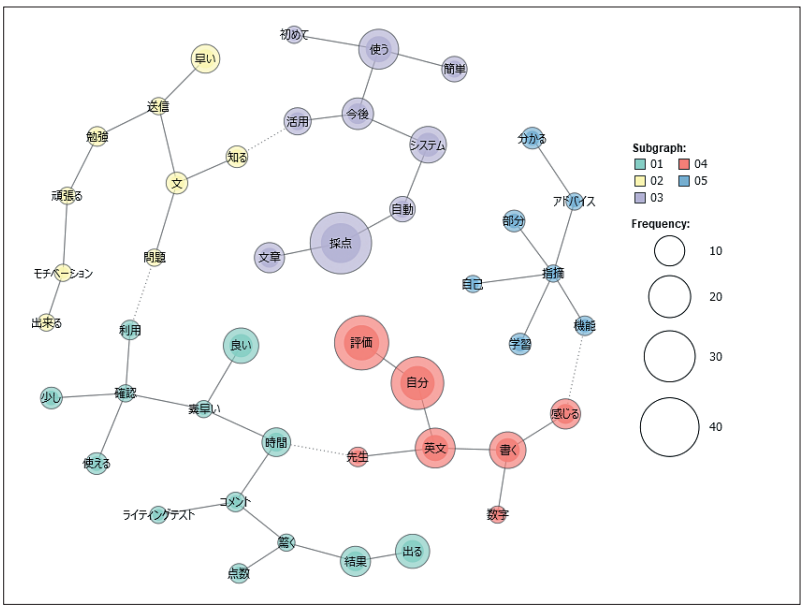

図６　自動採点システムを使った感想コメント

１つ目のサブグループでは，「良い」，「素早い」，「時間」，「コメント」，「驚く」，「結果」，「出る」，「点数」が示されていることから，良い点として素早い時間でコメント，点数，結果が出ることに驚いたという感想を待ったことがわかる．２つ目のサブグループには「早い」，「送信」，「勉強」，「頑張る」，「モチベーション」が示されており，送信するだけで結果がわかることで勉強を頑張るモチベーションが高まったことが推察される．３つ目のサブグループでは「採点」，「文章」，「自動」，「システム」，「今後」，「活用」，「使う」，「簡単」，「初めて」が示され，文章の自動採点システムは初めて使ったが簡単であり今後も活用したいと感じていることがわかる．４つ目のサブグループには「評価」，「自分」，「英文」，「書く」，「感じる」，「数字」が示され，自分の英文の評価が数字で書かれていることが印象的だったことが推察される．５つ目のサブグループには「指摘」，「部分」，「アドバイス」，「分かる」が示され，指摘された部分のアドバイスを望む意見が見られた．以下は各サブグループのコメントの抜粋である．

1. 読み込みに時間がかかると思っていたがコメントつきで約１秒で採点されたことに驚いた．将来，英文を添削する先生がいなくなってしまうのではと思った．
2. 送信してから結果が出るまでが早く，ストレスなく使えた．英語に対するモチベーションになり頑張ることができた．
3. 初めて使ってみて，今後も文章を作る機会があると思うので，このシステムを活用していきたいなと思いました．
4. 自分が書いたレベルがすぐに数字で示されてわかりやすかった．
5. 文章のアドバイスや文法の間違いを指摘してもらえないことは残念な部分だが，評価して

もらった後にとても自信がついたので良かった.

6. 3. 5. 修正した英文を入れてみた感想

質問 5「修正英文についての自動採点システムの評価についての感想」に対する学生の自由記述回答の総抽出語数は 2,386 語（249 文）であった．特徴を読み取りやすくするために「思う」という一般的な語を「品詞による語の取捨選択」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「評価」(52 回)，「変わる」(28 回)，「英文」(22 回)，「採点」(19 回)，「文」(19 回）であった.

「思う」を外した共起ネットワークを作成し，さらに「最小スパニングツリーだけを描画」を選択したところ，修正英文についての自動採点システムの評価についての感想に関して 7 つのサブグループが形成された（図 7）.

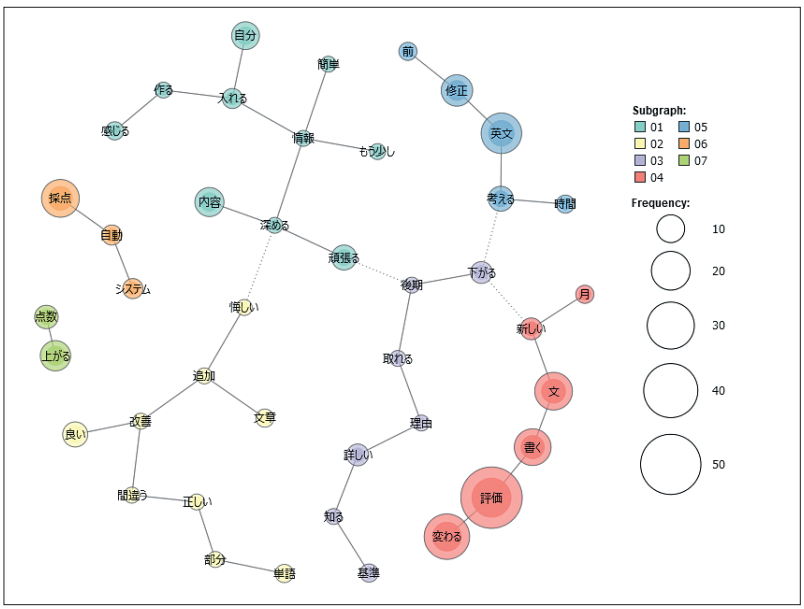

図 7　修正英文についての自動採点システムの評価についての感想コメント

1 つ目のサブグループでは，「内容」，「深める」，「頑張る」，「情報」，「簡単」，「もう少し」，「入れる」，「自分」，「作る」が示されていることから，英文を修正するにあたり内容を深めるように頑張ったことと情報をもう少し入れるように作ることを意識したことが分かる．2 つ目のサブグループには「追加」，「悔しい」，「文章」，「改善」，「良い」，「間違う」，「正しい」が示されており，文章を追加したあとの評価について改善されず悔しい思いと改善されて良かったという感想が現れている．3 つ目のサブグループでは「基準」，「知る」，「詳しい」，「理由」，「後期」，「取れる」が示され，「後期」は「頑張る」とも関連しており，修正英文のための評価基準と詳しい理由を知ることで後期最後のライティングテストでも良い評価を取れるよう頑張ろうという気持ちが読み取れる．4 つ目のサブグループには「変わる」，「評価」，「書く」，「文」，「新しい」が示さ

れ，新しく書いた文で評価がどのように変わったかについて関心が高いことが分かる．5 つ目のサブグループには「英文」,「修正」,「前」,「考える」,「時間」が示され，修正前の英文と見比べたいことや修正時に考える時間が少なかったと感じていることが推察される．6 つ目のサブグループには「採点」,「自動」,「システム」が示され，修正前と修正後の英文を入れたときの評価の変化から今回の自動採点システムの信頼性に関しておおむね好意的であることをうかがうことができる．7 つ目のサブグループには「点数」,「上がる」が示され，コメントに従って修正した英文の点数が上がった学生が一定数いることを読み取ることができる．以下は各サブグループのコメントの抜粋である．

1. 評価が変わらなかったのが悔しいですが，字数だけでなく内容もちゃんと考慮されているということだと思うので，内容を深められるように頑張ります．
2. 自分が思いつく限りの文章を追加したが評価が変わることはなく正直悔しい気持ちになりましたが，もっとネタを考えてやろうという励みにもなりました．
3. 内容を足してみたけど，評価が変わらなかったので，残り何を足したら 4 になるのかが詳しく知りたいと思った．
4. ・自分のことについて細かく書いたら評価が上がった．
   ・書き直すと評価が上がり，もっといい文を書きたいと思えるのでとても良いと思いました．
   ・自分が体験したことなどを入れてみたが，評価が変わらなかったのであとなにを入れたら良くなるのか知りたいです．
5. 修正前の英文と見比べられるようにしてほしい．
6. ・評価が変わらなかったので，自動採点システムは信用できる．
   ・細かく書いたら評価が上がった．
   ・7 月のライティングテストの自動採点が 4 だったので，新しく英文を作りましたが，新しく書いた英文の評価も 4 で嬉しかったです．
7. どうすれば点数が上がるのかアドバイスをもらえたのでそれにそって追加した時に点数が上がったので改善点を明確にわかるのはとても良いと思った．

7. 考察

2021 年度 3 回分のライティングテストの英文 512 件を機械学習の「教師あり学習」の入力データとして用いて自動採点システムを開発した．またそれを 2022 年度英語必修科目の授業内で学生に使用させた．この自動採点システムは，教育現場で限定されたタイミングで用いるための自動採点システムである．Amazon のクラウド上のサービスを利用し，技術者の協力により開発されたスモールスタートの AI モデルである．

自動採点システムで採点する項目は英文の「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 までの採点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，この自動採点システムでは

取り上げていない.

本システムを開発する前の予備調査では，2020 年度と 2021 年度の英文計 995 件と教師評価を入力データとして機械学習を行い，調査した．生データのみの実験の他に，生データをコピーして倍にしたデータセットを用いた場合，特定箇所を置換しデータ数を増やした場合の調査を行った．その結果，エッセイの内容や形式に制限を加え，表現を一定のパターンに限定することにより，少ないデータ量であっても特定条件下では有効なモデルが作成できる可能性があること，特定箇所を人工的に置換することはデータに偏りを生じさせることを確認し，本調査では人工的に表現を置換した英文を機械学習の入力データに用いないこととした.

本調査では 2021 年度ライティングテストの英文エッセイ 512 件をデータセットとして用いて自動採点システム（Model A）を開発し，それを 2022 年度後期の初回授業で学生に使用させた．学生たちには 2022 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 A に入力して採点結果を確認させた．リサーチクエスチョン（1)「学生のライティングテスト英文の教師による採点と自動採点システムによる採点の一致度はどの程度であるか」については，2022 年度の英文の計 124 件の自動採点が教師採点と一致した割合は 60.5％（75 件）で，初期モデルとしては予想を超える一致率であった．一致しない採点結果のうち 80.0％は自動採点と教師採点が 1 点差のものであった（49 件中 39 件）．1 点差をどう改善するかが次の課題となる．採点結果に大きな差があったサンプルを 6.2 節で紹介しているが，大きな差の理由は不明である.

今回開発した自動採点システム Model A は，2021 年度ライティングテストの英文データを用いている．今後は 2021 年度データに 2020 年度データを加えた Model B，さらに 2020 年度から 2022 年度までの 3 年分のデータを用いた Model C を開発して，一致率の変化を見ていきたい.

リサーチクエスチョン（2)「自動採点システムを利用した学生のアンケート回答から示唆されるものは何か」については，5 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，Model A の良い点として，多くの学生が自分の英文の評価得点が瞬時に，客観的に，明確に出ることを挙げ，英文改善の時間が短縮されることが評価されていることがわかった．このことによって，教師だけでなく学生にとっても素早い評価結果は学習の改善に大きな利点があることが確認された.

悪い点としては，誤りを正しく直すためのアドバイスや参考となる良い例文の紹介がコメントとして記載されることを求める意見が多く見られた．評価結果とコメントを表示する文字の見にくさに対する指摘もあり，次のシステム表示の改善を試みたい．また，文法エラーやスペリングエラーの具体的な指摘や訂正がない点に不満があることがわかった．ただし，この点に関してはこの自動評価システムでの採点は英文の「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，この自動採点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，次回以降使用前に「内容の質」に限定した評価点であることを周知していきたい.

Model A の改良点としては，自分の英文の問題点をより具体的に示し改善のためのアドバイスが提示されるシステムを望んでいることがわかった．次の Model B 開発においては，学生が

望むフィードバックに近づけるよう改善を試みたい．さらに，今回の Model A では評価の一致率が約 60％であるため採点結果に疑問を持った学生も多く，採点の正確さが改良点として指摘された．また，システムにログインする際の手間に改良の余地があることも分かった．これらも次回の Model B への課題としたい．

今回のアンケート項目で「これまでにライティングの自動評価システムを使ったことがあるか」という質問に対しては，全員が初めて使ったという回答であった．Model A を初めて使用した感想としては，英文を入力するだけで即座に評価の点数とコメントが出ることに驚いたこと，その場で結果がわかることで英文を作成するモチベーションが高まったことを指摘する声が多く，今後も活用したいとする意見も多数見られた．現在，日本の英語教育においてライティングを自動評価するシステムはまだ普及しておらず，今回のシステムは驚きをもって好意的に受け止められたようである．瞬時に採点結果が出ることが学生のライティングへの意欲を高めることが示唆され，その教育効果が期待できることが確認された．

学生が英文を修正して再度自動採点システムに入力した感想としては，点数と共に表示されるフィードバックに従って内容を深めたり情報を加えたりすることを意識したことで評価が上がった，または改善されず悔しい思いをしたという感想が見られた．こうした思いは，次回のライティングテストで良い評価を取れるよう頑張ろうという気持ちに繋がっているようである．今回の Model A では，点数とともにごく簡単なフィードバックが各評価につけられただけであったが，そこに含まれた「理由の詳細や具体例，印象的な情報，内容の深まり」といったコメントによって良い英文には何が必要かを自ら理解する学生も少なからずおり，端的なコメントでも効果をもたらしていることが確かめられた．

今回初めて試みた自動採点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．これまでのライティング研究においては「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 Model A の開発と実践を試みた．今後も引き続き学生の意見を反映させた Model B の開発に取り組んでいきたい．

8. 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人工知能の機械学習を行い，特定の授業での限定使用の自動採点システムを開発した．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採点負担を軽減し，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開発した．自動採点システムの採点項目は，特に採点が難しいとされる「内容の質」のみに限定した．今後も毎年度同一テーマでライティングテストを実施して機械学習用の入力データを増やしていき，自動採点システムの改良に努めたい．また自動採点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，成田康孝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 1 年前期に Integrated English a（1 年前期），Integrated English b（1 年後期）両方を開講している．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. Pigai とは北京词网科技有限公司（ベイジン・ツーワン・カージー・ヨウシャン・ゴンス ー běijīng cíwǎng kējì yǒuxiàn gōngsī）が提供するウェブベースの英作文支援ツールであり，中国語で批改网（ピーガイワン pīgǎiwǎng）と呼ばれるサービスの英語名称である．
3. CEFR-J とは，CEFR を日本人英語学習者のためにカスタマイズしたものである．
4. この業者はオンラインで英文添削サービスを提供している．世界中に英文添削講師を抱え，11 年の実績がある．
5. 以下は英語検定 2 級 2022 年度第 1 回ライティング試験で出題された問題である．
   以下の TOPIC について，あなたの意見とその理由を 2 つ書きなさい．POINTS は理由を書く際の参考となる観点を示したものです．ただし，これら以外の観点から理由を書いてもかまいません．語数の目安は 80 語～100 語です．
   TOPIC　Some people say that it is necessary for people to go to important historical sites in order to understand history better. Do you agree with this opinion?
   POINTS　Experience, Motivation, Technology
6. 以下はリライトされたサンプルである．
   例（2021 年 4 月）：
   リライト前（順番のディスコースマーカーがないので 0 評価，42 語，Level 0）
   The place I want to visit most is Korea. There are three reasons.
   Because I like Korea idol very much.
   Korea food is very delicious.
   oman is very buautiful.
   Japan and Korea is near.
   I think go to Korea very much.

   リライト後（順番のディスコースマーカーを追加，45 語，Level 1）
   The place I want to visit most is Korea. There are three reasons.
   First, I like Korean idols very much.
   Second, Korean food is very delicious.
   Third, Korean women is very beautiful.
   Japan and Korea is near.
   I want to go to Korea very much.

   例 2（2021 年 7 月）：
   リライト前（文の途中で終わっているので 0 評価，75 語，Level 0）
   The place I want to visit most is Disneyland. There are three reasons.
   First, I have only been Disneyland 3 times in my childhood. So I want to go there to see the difference from when I was a child.
   Second, I saw Disney movie lately and I though If I go to Disneyland, I can meet to the Disney character.
   I heard that there is a "Greeting", so I definitely want to go.
   Finaliy,

   リライト後（3 つ目の理由を追加，86 語，Level 3）
   The place I want to visit most is Disneyland. There are three reasons.
   First, I have only been to Disneyland 3 times in my childhood. So I want to go there to see the difference from when I was a child.
   Second, I saw a Disney movie lately and I thought if I go to Disneyland, I can meet the Disney character.
   I heard that there is a "Greeting", so I definitely want to go.
   Finally, I want to ride the attractions. They must be fun.

   例 3（2022 年 1 月）：
   リライト前（最初から 2 つの理由としているので 0 評価，77 語，Level 0）

The place I want to visit most is Guam.
There are two reasons.
First,I love nature, especially the ocean.
The sea of guam is clear.
We can enjoy marine sports in the sea.
For example, we can enjoy jet skiing etc...
Second,I want to enjoy shopping.
Because, in Guam, we can get brand products cheaper than in Japan.
It's very exciting for me, who loves cosmetics and clothes.
From the above, I wanted to go to Guam.

リライト後（3 つの理由に変更，92 語，Level 3）
The place I want to visit most is Guam.
There are three reasons.
First, I love nature, especially the ocean.
The sea of Guam is clear.
We can enjoy marine sports in the sea.
For example, we can enjoy jet skiing etc...
Second, I want to enjoy shopping because in Guam, we can get brand products cheaper than in Japan.
It's very exciting for me to go shopping, because I love cosmetics and clothes. And finally, I want to travel with my friend.
For these reasons, I wanted to go to Guam.

7. 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．
8. 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．
9. 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（2020a）「自動採点研究とは？」石井雄隆＆近藤悠介（編）『英語教育における自動採点…現状と課題』（pp. 1–15）．ひつじ書房．
石井雄隆・近藤悠介（2020b）「教室における指導と自動採点」石井雄隆＆近藤悠介（編）『英語教育における自動採点…現状と課題』（pp. 117–130）．ひつじ書房．
石岡恒憲（2020）「自動採点研究のこれから」石井雄隆＆近藤悠介（編）『英語教育における自動採点…現状と課題』（pp. 131–156）．ひつじ書房．
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．
小林雄一郎（2020）「学習者コーパス研究と自動採点」石井雄隆＆近藤悠介（編）『英語教育における自動採点…現状と課題』（pp. 73–93）．ひつじ書房．
近藤悠介・石井雄隆（2017）．英語学習者の発話自動採点システムの開発と英語教育プログラムへの導入可能性の検討．*Language education & technology, 54*, 23–40.
三田薫・霜田敦子（2020）学生の英文ライティング力向上の分析―Fluency が伸びた学生の日本語の干渉によるエラーと表現力の変化．*Jissen English Communication, 50*, 6–33.
三田薫・霜田敦子（2021a）学生の習熟度別 英文ライティング力向上の分析―弱点克服の重点的指導によるライティングの変化―．*実践女子大学短期大学部紀要 =Jissen Women's Junior College Review, 42*, 63–83.
三田薫・霜田敦子（2021b）学生の英文ライティング力向上の分析 その 2: 文法・構造・論理の重点的指導によるライティングの習熟度別変化．*Jissen English communication, 51*, 14–46.
三田薫・霜田敦子（2022a）英語初級学習者のパラグラフ・ライティング評価基準の確立を目指して．*実践女子大学短期大学部紀要 =Jissen Women's Junior College Review, 43*, 65–83.
三田薫・霜田敦子（2022b）学生の英文ライティング力向上の分析 その 3: 文法・構造・論理・内容の質の重点的指導によるライティングの習熟度別変化．*Jissen English communication, 52*, 13–48.
中谷安男（2019）『英文エッセイの自動レベル判定システムと手動採点結果の比較検証：CEFR-J ライティング・テストタスク構築ための予備調査』法政大学経済学部学会 *経済志林，87*（1, 2）, 21–50.
永田亮（2020）「深層学習に基づいたエッセイの自動採点」石井雄隆＆近藤悠介（編）『英語教育における自動採点…現状と課題』（pp. 95–115）．ひつじ書房．
日本英語検定協会（2018）「AI による自動採点実証研究で有意な成果…2019 年度から英検に順次本格導入予定…」https://www.eiken.or.jp/eiken/info/2018/pdf/20181017_pressrelease_aisaiten.pdf

2022 年 8 月 18 日閲覧.
小田登志子 (2017). 迫り来るライティング時代に対応する英作文支援ツールとは: 中国 Pigai の事例報告: 研究ノート.
https://repository.tku.ac.jp/dspace/bitstream/11150/10923/1/jinbun140-10.pdf
2022 年 8 月 28 日閲覧.
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.
Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic Text Scoring Using Neural Networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 715–725.
Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. *Proceedings of the Thirty-sixth Annual Meeting of the Association for Computational Linguistics*, 206–210.
Ke, Z., Inamdar, H., Lin, H., & Ng, V. (2019). Give Me More Feedback II: Annotating Thesis Strength and Related Attributes in Student Essays. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3994–4004.
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.
Li, Z., Link, S., Ma, H., Yang, H., & Hedleheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*, 66–78.
Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard". *Applied Measurement in Education, 28* (2), 130–142.
Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: a comparative study. *Journal of Technology, Learning, and Assessment 6* (2), 4–28.
Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *International Journal of Language Testing, 2* (1), 59–92.
Yoon, S-Y., Evanini, K., & Zechner, K. (2011). Non-scorable Response Detection for Automated Speaking Proficiency Assessment. *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, 152–160.
Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*, 883–895.
Zribi, R. & Smaoui, C. (2021) Automated versus Human Essay Scoring: A Comparative Study. International Journal of *Information Technology and Language Studies* (*IJITLS*), 62–71.