博士論文

ルーブリックに基づくレポート自動採点支援システムの研究

D2016SC001　山本　恵

指導教員　河野　浩之

2019 年 2 月

南山大学 理工学研究科 機械電子制御工学専攻 博士後期課程

Studies on Automated Essay Scoring Support System Based on Rubric

D2016SC001　Megumi Yamamoto

Supervisor　Hiroyuki Kawano

February 2019

Graduate Program of Mechatronics

Graduate School of Science and Engineering

Nanzan University

要約

本研究の目的は，大学の基礎教育の現場において教員・学生双方に役立つレポート自動採点支援システムの構築である．近年，自由記述文やレポートによる評価の重要性が論じられ，採点の厳正化や効率化が課題となっている．これらを解決するための一つとして自動採点システムの研究が進められているが，大規模試験の合否判定を目的とした研究が中心で，授業改善や学生指導を中心とした大学の教育現場で活用するシステムの研究はほとんど見当たらない．

これらの課題を解決するための，レポート自動採点支援システムを構築している．システムは学習管理システム（Learning Management System，LMS) 上に構築し，通常の授業運営の中で教員・学生ともに主体的な利用を可能とすることで，効率的な利用や負担軽減をはかっている．具体的には，梅村らとの共同研究で開発している Moodle のテキストマイニング・プラグイン TeMP[1] を拡張し，本採点支援システムを構築する．Moodle はシステムのポータルとしての役割を持ち，データ収集や採点結果のフィードバックと，採点処理を行う外部サーバとのデータ授受を行う．これらは Moodle と同様に PHP を用いて開発している．Moodle を通して得られた記述文は，データ解析用サーバ（クラウド型アプリケーション）で処理され，教員に対して評点とともに論作文指導に有用な言語統計情報を提供する．学生に対しては，文書検査サーバ RedPen により文章校正を行い，レポートの文章校正結果をメッセージ形式でフィードバックする．データ解析用サーバは，MeCab（形態素解析器）による処理， R（分類，重回帰，その他の計算）による採点処理を行っている．文書検査サーバはオープンソースの RedPen でありを利用しており，文法や構文エラーをチェックする．

研究の特徴は，以下の 2 点である．

(1) 評価の厳正化を保つために，レポート採点用ルーブリックを基盤とした採点支援モデルを提案する．

(2) 語彙力および記述内容の採点にあたり，網羅性の高い語彙レベル辞書構築モデルを提案し，日本語 Wikipedia などの大規模コーパスから作成した辞書を利用する．

(1) のルーブリック作成では，アメリカの Value Rubric[2] や国内の汎用的なルーブリック

を参考に，5 評価観点と採点支援技術を踏まえてそれらを細分化した 25 評価項目を提案している．本研究ではこのうちの 12 評価項目を自動採点し，その結果から重回帰モデルにより 2 つの評価観点を予測する．さらに，実際のレポートの評価結果を学習データとしてサポートベクターマシン（Support Vector Machine，SVM) で分類器を作成し，総合成績を 5 段階のレターグレードで分類する．採点精度は手動採点との相関を目安としており，総合成績を予測する分類では 53.6%である．実際の教育現場での試用において，授業改善を示唆する有用な結果が得られている．

(2) では，採点精度向上に向けて，レポートの語彙水準評価で用いる語彙レベル辞書の構築モデルを提案している．具体的には，大規模コースへのトピックモデル (LDA) 適用により算出された出現確率をもとに，単語難易度を計算し，さらに TF-IDF による値の補完を行い，記述文の語彙水準を測るための語彙レベル辞書作成方法を提案している．大学生が使用する語彙を豊富に含む日本語 Wikipedia コーパスを利用して辞書を構築し採点した結果，採点精度が 4.9%向上している．

大学の基礎教育のレポートでは，文章作成スキルや語彙力が共通の評価項目となる．これらの表層的な特徴量からレポートの総合評価を求めた結果，ある程度の精度が確認できたこと，教育現場での試用において，教員学生双方に役立つ結果が得られたことから，教員・学生双方を支援する自動採点システムとして利用可能である．精度向上のための語彙レベル辞書構築手法は，元となるコーパスに依存しないため，専門教育でのレポート評価の採点に応用することが可能である．今後は，近年の自然言語処理技術の進歩を視野に，論理的内容の評価に踏み込んだ採点に展開することで，さらに高い採点精度を期待できる．

Abstract

The aim of this study is to construct the automated essay scoring support system useful for both teachers and students at the university's basic education classes. In recent years, it is discussed the necessity of active learning and the strict evaluation. It is difficult to evaluate the results of written essays in such classes in the conventional examinations, therefore performance evaluation like essays, papers, presentations are adopted. Teachers are required to support students for writing and score their essays strictly. The most currently researches on the automated essay scoring system are aiming to pass / fail judgment of large-scale examination. We can barely find any research on systems to utilize on site focusing on improvement of grades and education for university students.

In order to solve these problems, we propose a rubric for human scoring and introduce it to the automated essay scoring system and construct the essay scoring support system.

Our system is constructed on the Learning Management System (LMS), making it possible for both teachers and students to use themselves in the course of ordinary lessons, thereby making efficient use and reducing the burden. Specifically, we extend the Moodle plug-in TeMP[1], which is the text mining system developed by collaborative research with Umemura et al. Moodle has a role as a system portal, and collects data, feeds back the scoring result, and exchanges data with an external server that performs scoring processing. The essays obtained through Moodle are processed by the data analysis server (cloud type application). Teachers are supplied the score and linguistic statistical information. The data analysis server performs scoring process by MeCab (morphological analyzer) and R (SVM, multiple regression, other calculation). It is used the open source RedPen server as the document inspection server, which checks grammar and syntax errors.

The characteristics of this study are the following two points:

(1) The scoring support model based on rubric to maintain strict evaluation

(2) The highly comprehensive vocabulary level dictionary construction model

Firstly, the rubric is developed with reference to the Written Communication VALUE Rubric[2] in AAC&U of the United States and common rubric in Japanese university. This has five evaluation viewpoints and 25 evaluation items. We evaluate 12 evaluation items among them automatically, and predict two evaluation viewpoints from the result by multiple regression model. Furthermore, the classifier is created by support vector machines from scored essays as learning data. The new essays are classified by it to the final scores. Scoring accuracy is based on the correlation with human scoring, and it is 53.6 % in classification that predicts a total score. In the actual trial at the educational site, useful results suggesting class improvement are obtained.

Secondly, we propose the construction model of the vocabulary level dictionary used in the vocabulary level evaluation of essays to better the scoring accuracy. Specifically, the word difficulty level is calculated based on the appearance probability calculated by applying the topic model using the LDA to the large-scale corpus, and the value is further complemented by the TF-IDF.

In the basic education of the university, sentence making skills and vocabulary skills are common evaluation items. As a result of a comprehensive evaluation of the report from these feature quantities, it was confirmed that some degree of accuracy was confirmed from the trial at the educational site, and useful results were obtained for both faculty and students. This system can be used as the automated essay scoring support system. In addition, from the viewpoint of progress of natural language processing technology in recent years, further higher scoring accuracy can be expected by developing to grades that have stepped into the evaluation of logical contents.

目 次

表 目 次

図 目 次

第1章　序論

本章では，研究の目的やシステム構築アプローチについて言及する．1.1 節で教育現場における課題と研究の目的を，1.2 節でレポート自動採点システム構築のアプローチを，1.3 節で本研究の特徴を，1.4 節で本稿の構成について述べる．

1.1　レポート採点の課題と研究の目的

近年，大学教育では，アクティブラーニング（能動的学習）の必要性や，それらの授業でのパフォーマンス評価におけるルーブリックの重要性が議論されている．教員による一方向的な講義形式から，学修者の能動的な参加を取り入れた授業形態（体験学習，ディスカッション，プレゼンテーションなど）が研究・実践されており，このような授業での学修成果を，従来のテスト形式で評価することは困難である．そのため，レポートや論文，プレゼンテーション，作品，協調活動など，いわゆるパフォーマンスによる評価方法が採用される．特にレポートは，学習者の知識はもとより，思考力や問題解決能力など多くの習熟度を測ることができ，今後ますます有用な評価方法として利用されると考えられる．こうしたレポート評価では，採点者（評価者）による評価結果のばらつき，同一採点者内での評価の偏り，採点者の時間的負担など，様々な問題がある．したがって学生のレポート作成指導や，教員の採点の厳正化，負担軽減が希求の課題である．さらにレポート評価は，成績評価だけでなく，学生の論作文能力の育成も目的の一つであるが，実際には教員が十分な指導時間を確保できない，あるいは，指導のための個別資料を効率よく得ることが困難などの問題がある．

これらを解決するための方法の１つとして，自動採点システム（Automated Essay Scoring System，AESS）の研究と導入が進められ，アメリカではすでに商用化されているものもある．日本でも大学センター試験で記述文を取り扱うことの重要性が指摘され，自動採点の研究が進められているが，広く運用されているレベルのものはまだない．また国内外とも，大規模試験の採点における自動採点システムの開発が中心である．そこで本研究では，教育現場での活用を目的として開発する．

レポートを採点する場合，多くの採点者はチェックリストや採点の指標（いわゆるルーブリック）を定めて評価の厳正化を保つ努力をしている．自動採点システムを構築するにあたり，こうしたルーブリックを基準に採点のアルゴリズムを設計することは自然である．また

学習管理システム（Learning Management System，LMS）上に構築することで，データの収集から採点結果を得るまでの教員の作業が簡略化できる，学生自ら投稿し文章校正結果を得ることで，論作文スキル向上に向けた主体的な学習を行うことができるなどのメリットがある．

以上より，大学の教育現場において，レポート採点に関わる教員の諸問題（採点負担，評価の揺らぎ）を解決し，学生のレポート作成能力育成を支援する，ルーブリックに基づく自動採点支援システムを Moodle プラグインとして開発する．

1.2 AESS 構築のアプローチ

構築する自動採点支援システムの中枢を担うのは，レポートなどの自由記述文を評価し採点する自動採点システムである．自由記述文はテキスト集合であり，これらを採点すなわち評価するには，テキストマイニング技術が不可欠である．

ローネンらはテキストマイニング技術について「（１）発見のためのマイニング」「（２）検索のためのマイニング」「（３）情報分析のためのマイニング」の 3 つの流れを示している [3]. 本研究は（３）に相当し，テキスト集合 (レポートなどの自由記述文データ) に散在する情報を分析・解釈し，知見（評価値）を得るためのシステムである．したがって，AESS をテキストマイニングシステムとして位置付け，構築する．

一般的なテキストマイニングシステムの処理に AESS をあてはめ，図 1.2.1 に示す．

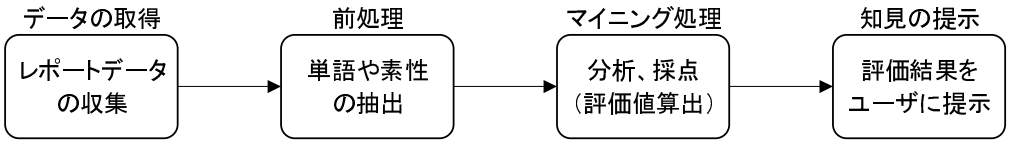

図 1.2.1: AESS 処理手順

AESS の研究課題は，図 1.2.1 の各フェーズで以下のように整理できる．

1) データ取得時：解答データ入力の多様化（音声や手書き文字の認識）への対応

2) 前処理：形態素解析，単語の抽出やアノテーションの付加

3) マイニング処理：採点アルゴリズムの研究（人間の採点結果と比較して精度を示す）

4) 知見の提示：採点結果の提示方法の研究（教員への有用な情報となる情報の提示，学生の論作文能力育成に向けた文章校正のフィードバックや指導の研究）

本研究では，3) と 4) を扱い，特に 3) に重点を置く．なお 1) については，LMS の授業サイトへのレポート（テキストデータ）提出を前提に開発している．2) については，既存の形態素解析システム MeCab を利用する．3) については，統計的手法で行う．4) は，教育学の理論的なアプローチではなく，構築した AES 支援システムのプロトタイプを利用して，教育現場で試行しながら改善を目指すというアプローチで行う．

以下，3) について補足する．アプローチの手法としては，論理的推論および確率・統計モデルからの推論がある．

a) 論理的推論：自然言語処理の分野．文章 1 つ 1 つの意味を解析する

b) 確率・統計モデル：言語資源（特に記述文や書き言葉などのコーパス）や，すでに獲得している過去のレポートデータのマイニング結果を利用する

本研究では，b) の統計的手法で，精度の高いシステムの確立をめざしている．現在もっとも AESS の開発が進んでいるアメリカの先行研究をみると，b) あるいは双方を取り入れる研究が多い．大量のデータを入手し処理できる技術や環境が整ってきたことが，一つの理由と考えられる．現在，全て機械に任せる完全な自動採点システムが完成されたわけではなく，精度の高いシステムの確立は未だ困難である．特に a) の記述文の意味解析において，手動採点と同じ成果（精度 100%）を出すレベルには至っていない．また実際の教育現場では，内容については教員がレポートを読み込んで判断すべきであり，自動採点の導入には賛否両論がある．そこで本研究では，あくまでも支援システムとしての役割を担うべきであるとして，自動と手動のハイブリッド型である自動採点支援システム（Automated Essay Scoring Support System）を開発している．以降では，先行研究の自動採点システムを AESS，本研究の自動採点支援システムを AES 支援システムと表記する．

1.3 研究の特徴

本研究の特徴は，以下の 2 点である．

1) レポート採点用ルーブリックを基盤に自動採点システムを設計

ルーブリックに基づく採点および結果のフィードバックは，教員・学生双方にとって授業目標の達成度の共通認識につながるとされ，近年，その導入が推奨されている．そこで，

レポート採点用ルーブリックを作成し，これを基盤に採点アルゴリズムを設計する．また，授業担当者がレポートを採点する際に，自動採点結果をセカンドオピニオンとして参照する，あるいは文章作成スキル部分の評価に自動採点結果を利用することで，評価の厳正化や時間的負担軽減を図る．

2) 採点精度向上のための語彙レベル辞書を構築する
レポートの語彙水準採点にあたり，Wikipedia，新聞，雑誌，論文などの大規模コーパスを利用した語彙レベル辞書の構築モデルを提案する．合否を判定する試験の自動採点では，正解データを学習させるなど，事前に正答例やスコアが高い採点済みデータを投入し，各種の教師あり学習により，比較的高い精度で採点可能である．しかし，教育現場で採点するレポートは，科目や教員ごとに設問や目標達成度が異なり，多数の正答例を準備することが困難である．そこで教員が影響を受ける可能性が高い語彙水準の算出は重要であると考える．本研究では，大規模なコーパスに潜在ディリクレ配分法（Latent Dirichlet Allocation，LDA）を適用し，単語出現確率から語彙の難易度を計算し，語彙レベル辞書を構築する手法を提案する．時代とともに使われる単語は変化する．提案したモデルは，もととなるコーパスに依存しないため，相応しいコーパスを選択し構築し直すことが可能である．

1.4 本稿の構成

本稿は 8 章で構成される．本章で研究の目的を述べた後に，第 2 章では，AESS やコーパスなどに関する先行研究を概観し，要素・技術について整理する．第 3 章では，AES 支援システムの採点モデルの全体像とルーブリックの必要性および提案，第 4 章では，AES 支援システムのアーキテクチャと機能や採点アルゴリズムについて述べる．第 5 章で，AES 支援システムを用いた実験結果を報告し，自動採点対象となる評価観点・評価項目およびシステムの基盤となる自動採点部分の妥当性を確認する．第 6 章では，教員および学生それぞれの利用者からみた改善点と改善策について考察する．第 7 章では，実験の結果から確認された精度の問題と改善策を報告し，第 8 章でまとめと今後の課題を述べ，むすびとする．

第2章　先行研究

本研究に関連する先行研究として，2.1 節で AESS，2.2 節でルーブリック，2.3 節でコーパスからの辞書作成について紹介する．

2.1　AESS 関連研究

本節では，AESS の研究の歴史や動向を調査し，現在実用化されているシステムがどのような要素技術を使っているのかを比較しながら整理する．2.1.1 では，AESS に関わる技術の発展と AESS 研究の歴史を概観する．2.1.2 では，代表的な自動採点システムの比較を行い，2.1.3 で日本国内における研究状況を述べる．2.1.4 で，AESS にかかわる要素・技術，特に採点手法を整理する．

2.1.1　AESS に関わる要素技術と研究の歴史

AESS の研究は 1960 年代のアメリカで，採点者の負担軽減を目的として始まった [4]．ハードウェア，ソフトウェア，自然言語処理，情報検索技術など，様々な要素技術の発展の影響を受けながら，現在も研究・開発が続いている．Mark ら (2013) は，AESS の発展に寄与した要因として，ワードプロセッシング，インターネット，自然言語処理（Natural Language Processing，NLP）の 3 つの進歩と発展をあげている [5]．石岡（2004）は，1980 年代の作文ツール（Writers Workbench，WWB）や，特に日本における文章校正支援ツールの存在を上げている [6]．これらに加え，ハードウェア（CPU の処理速度），人工知能（Artificial Intelligence，AI），データの蓄積とビッグデータ管理システム，テキストマイニング，コーパスなども重要な要素技術である．黒橋 (2015) は自然言語処理の歴史を，黎明期→忍耐期→発展期としてまとめている [7]．特に発展期の，ビッグデータ，テキストマイニング，コーパスの急速な進展により，AESS の研究はより活発化し，現在に至る．以下に動向を概観する．

1940 年代半ばのコンピューターが生まれた時代は黒橋が言うところの，自然言語処理の黎明期である．AESS の研究はなされておらず，暗号解読などの軍事目的で行われていた．またロシア人が初の人工衛星打ち上げに成功するなどを機に，ロシア語から英語への機械翻訳に関心が高まった．コンピューターによりテキストデータを蓄積し，検索する試みも始まり，1960 年代にコーネル大学で情報検索システム SMART が開発され，ベクトル空間モデルなど

の概念が提案されている．また 1956 年のダートマス会議で「人工知能」という言葉が初めて使われ，コンピューターの言語理解や質問応答など，自然言語に関する研究にも関心が持たれた．しかし，コンピューターの処理能力が不十分で，十分な実証研究が困難な時代であった．

1960 年代半ばから 1990 年頃は自然言語処理の忍耐期とされる． コンピューターの処理速度が上がったものの，自然言語処理には十分ではなく，開発環境が整わなかった時代と言える．研究が進むにつれ問題の難しさが認識され，アメリカでは機械翻訳への研究費が制限されていた．AESS の研究については萌芽期と言える．Page は，1966 年，アメリカの教育系専門誌 Phi Delta Kappan に”The imminence of grading essays by computer” と題して，コンピューターによる AESS の必要性を訴えている [8]．論作文能力育成には文章を多く書くことが必要と考えていたが，採点評価の負担が障害となり課題を出すことができない現状を改善しようとしたためである． しかしながら，1960 年代はテキスト入力としてパンチカードが主流で，ワープロソフトも 10 年待たなければならず，人間の評価者に機械が取って代わることへの反対意見など，ハードウェア，ソフトウェア，世論を含む様々な点で，Page の AESS のアイデアを実現するには困難な時代であった．一方，カナダや日本では英語への翻訳システムが，またヨーロッパでは多言語機械翻訳システムの研究開発が盛んになった．1960 年代後半からファイルのデータベース化がすすめられ，1970 年に関係モデルの概念が提案された [9]．この頃オンライン文献データベースのサービスが開始されている．また 1967 年にはじめて言語コーパス Brown Corpus が発表されている [10]．しかしながら機械翻訳などの知的処理ではまだ精度が足りず，人間が整合性をとる段階であった．1980 年代の終わりには，世界最大の試験機関であるアメリカの ETS が，e-rater を開発している [11]．e-rater は，採点のための多数の特徴量から重回帰モデルにより得点を求める．各エッセイについて 2〜3 人の専門の評価者も採点し，自動採点結果との差異が大きい場合，具体的には 6 点満点で 2 点の差があると，確認し最終的な評価を決定している．このように採点のすべてを機械に任せるという完全な自動化ではなく，人間の採点結果と照合している．

1990 年頃からは，AESS の研究に役立つ技術が進展し，自然言語処理の発展期に入る．1990 年に Berners-Lee により World Wide Web が提唱され，インターネットが普及し社会基盤となった．さらにインターネットを経由して自動収集される大規模なコーパスなど，言語資源の収集が容易になってきた．機械翻訳についても対訳コーパスに基づく翻訳が提案され大きな進展がみられた．テキストマイニング技術と言語処理技術の交流が活発化しさらに研究が進展している．またコンピューターの処理能力の向上とともに，機械学習，自然言語処理，情報検索，ナレッジマネージメントなど様々な理論や技術が発展した．潜在意味解析（Latent

Semantic Analysis，LDA）やコサイン類似度によるスコア計算，ベイズ理論，ルール発見による分類など，内容の評価に関する採点アルゴリズムが提案されるようになった．2000 年代は，採点そのものよりも解答データ入力の多様化（音声や手書き文字の認識）への対応や，学生の文章作成能力育成に向けた文章校正のフィードバックや指導の研究にシフトしつつあった [12]．このようなシフトは，全て機械に任せる完全な自動評価システムが完成されたわけではない．精度の高いシステムの確立が困難であり，次のステップに向かう足踏み状態の中で，自動採点システムのもう 1 つの目的である文章作成能力の育成に注力せざるを得ない状況だったのである．また，英語圏で発展した自動採点システムは，自動翻訳システムの発展とともに，他言語への対応のための開発が進められている．同時に，韓国や日本など大規模な公的試験が実施される国では，採点のコスト削減や採点者の負担軽減を目的として，アメリカで発展した自動採点システムを参考にしながら開発が進められるようになった．

日本における自動採点システムの発展は，その前身とも言える文章校正支援システムに依るところが大きい．1980 年代には出版業界を中心に，文章校正支援システムが開発され，実用化されている．代表的なシステムは，VOICE-TWIN[1]，St.WORDS[2]，および FleCS[3]である．出版分野では語の使用ルールが明確であるため，文章校正支援システムは長く利用されている．1990 年代に入り，こうした文章校正支援システムと自然言語処理などの技術の進歩とともに，記述文の評価の測定に役立てる試みがなされるようになった．2003 年には石岡らにより，当時唯一の日本語小論文の自動評価システム Jess が開発・発表されている [13].

2010 年頃からは，新たな採点アルゴリズムの提案よりも，コンピューターの処理能力の向上や，ビッグデータ，コーパスの利用が促進され，機械学習の様々なアルゴリズムを比較・検証することができるようになった．また各国で共通試験・統一試験などの大規模な試験の採点需要が増えた．特にアメリカでは AESS 開発がビジネスとしてり立ち，競争化している．2012 年にはヒューレット財団（The William and Flora Hewlett Foundation）がスポンサーとなって，AESS のコンペティションが実施され，154 チームが参加した[4]．Web 上でエントリーし，約 3 か月間で 7〜8 年生のエッセイ 8 セット（平均 150〜550 ワード）を採点し，性能を競うものである．e-rater， IEA，IntelliMetric など，アメリカの代表的な AESS が招待された．近年は，短答式試験の自動採点システムの開発にシフトしている．アメリカでは c-rater（ETS）や CRASE（Pacific Metric 社）がある．日本では石岡らが中心となり，センター試験利用に

[1]日経新聞が利用．NTT が開発した REVISE を母体とする
[2]講談社が利用．COMET を母体とする
[3]産経新聞が利用
[4]https://www.kaggle.com/c/asap-aes

向けて JS4 の開発が進められている [14]．短答式試験は望ましい正解文があるため，回答文と同義であるかを判定する含意関係認識技術が用いられる．

2.1.2 代表的な AESS の比較

アメリカでは，PEG，e-rater(現在は e-rater V.2)，IntelliMetric，AutoScore，LightSIDE，Bookette，Lexile Writing Analyzer，CRASE で，9 割以上の市場を占めていると言われる．石岡は文献 [15] および [16] で,「エッセイ評価システムの比較」として代表的な自動採点システムを表にまとめている．これらを引用しさらに大規模な試験での利用を目的に開発された他の AESS や実際に運用されている MOOCs の採点システムなどを追記し，表 2.1.1 を作成した．比較項目については「システム利用目的」,「商用」に対応しているかどうか（○は対応、非は非対応を意味する）,「ルーブリック」を自動採点に導入しているかどうかを追記した．また，開発年または関係論文が公開された年がわかるものについては開発者欄に付記してある．表の記載順序は概ね開発時期の昇順である．不明なものについては順序不問で後部にまとめてある．表に記載した順序にしたがい，各システムの概要を述べる．なお，引用以外の部分の記載内容については商用のシステムでは Web 上で公開されている情報を，その他開発関係者や研究者の論文を参考にした．また公表されていない，あるいは確認できなかったものについては「-」で示した．

表 2.1.1: AESS の比較

採点システム	開発者 (開発または論文等公開年)	システム利用等	評価基準	ルーブリック	商用	手法	制限，特徴など
PEG	Measurement Inc. (1973)	商用のオンラインテスト (MIST) の採点ほか商用（MI が販売・管理）	構造／組織化／形式／技巧／独創性．当初は特性（流暢さ，言葉遣い，文法，構造など）を反映する 300 以上の尺度を計算	無	○	現在は機械学習による分類	スコアは連続値
e-rater V.2	ETS (1988〜2004)	TOEFL（reading comprehension tests）の手書き答案の自動採点，アメリカ経営大学院入学試験 GMAT の小論文の採点 (〜2005)	構造／組織化／内容．一般的な作文能力を採点するにあたり，12 の特徴量 (評価指標) を計算．論題によらず固定．	無	○	重回帰モデル	スコアは連続値
IEA	Pearson Education (1999)	商用（アメリカのテスト機関 PKT 社が販売）	3 つの観点（内容/文体/技巧）と総合スコア A-D，F に分類	無	○	LSA，コサイン類似度	スコアは連続値

次ページに続く

前ページからの続き

採点システム	開発者 (開発または論文等公開年)	システム利用目的等	評価基準	ルーブリック	商用	手法	制限，特徴など
BETSY	メリーランド大学 Rudner ら (2002)	研究目的（フリーでダウンロード可能）	4〜6 段階のクラスに分類	無	非	ベイズ理論，ユニグラム言語モデル	700 ワード以上のエッセイには学習が十分でない．スコアはカテゴリーで示す
IntelliMetric	Vantage Learning (2003)	ペンシルバニア州の司法試験や大学入試試験の論述式問題採点，GMAT(2007 より)	一貫性／内容／構成／文章の複雑さ／アメリカ英語への適用	無	○	模範解答により人間の採点者の採点ルールの判断を推定	論題ごとに大量のデータが必要．スコアはカテゴリーで示す
Jess	Ishioka & Kameda (2003)	大学入試センター試験の記述式問題の採点	修辞／論理構成／内容．e-rater の採点を基盤に，日本語の小論文の一般的な採点として評価項目を設定	無	非	外れ値検出 & LSI	プロのライターの文書を利用．科学技術分野に弱い
Bookette	CTB/McGraw-Hill (2005)	2009 年より大規模試験で使用	構造・文法・意味・技巧	無	○	ニューラルネット．専門家のスコアをモデル化	90 の特徴量

次ページに続く

前ページからの続き

採点システム	開発者 (開発または論文等公開年)	システム利用目的等	評価基準	ルーブリック	商用	手法	制限，特徴など
LightSIDE	カーネギーメロン大学 (2009)	オープンソースのテキストマイニングツール	内容や文体	無	非	教師あり学習による分類	Weka を利用．初心者向け GUI 環境
AI grading	edX （2013）	オンライン公開講座 MOOCs の採点システム	読みやすさ，文字数やワード数	一部利用	○	NLP と機械学習	人間が採点した 100 の小論文の採点パターンを利用．他の論題には不向き
KASS	Eun-Seo Jang (2014)	韓国の共通試験 CSAT 等での利用を視野に開発	記述内容	無	非	LSI やコサイン類似度（模範解答との類似度）	基本的に短答式解答への対応
AutoScore	American Institutes for Reserch	州や地域での教育現場での利用	意味概念/段落間のつながりを示す意味的尺度／語彙，文法	無	-	統計的手法	教師データをもとに論題ごとに採点基準を作成
Lexile Writing Analyzer	MetaMetrics	教育関係製品の一つとして開発	語彙使用の多様性など個人の基礎的な文章表現能力を予測	無	○	統計的手法	学習データ不要
CRASE	Pacific Metrics	次世代評価システムの開発	アイデア／文章の流暢さ／組織化／態／語彙選択／慣習／プレゼンテーションのうまさ	無	○	機械学習＋統計（ベイズアプローチ）	Java ベースの Web アプリケーション

以上

Project Essay Grade（PEG）は，1966 年，Page によって提案された [8]．文の長さや句の数など 30 の特徴量を用いた重回帰モデルを用いていた．大学レベルの短いエッセイでは良好に機能したようであるが，表層的な特徴量であるため，ハイスコアを得るためのトリックが可能であるとの批判があった．その後，1993 年に改訂され，いくつかの構文解析器や辞書が追加されている．2002 年には Measurement Incorporated(MI) が Page らから権利を取得してメンテナンスを行っており，商用で実用化されている．特性（流暢さ，言葉遣い，文法，構造など）を反映する 300 以上の尺度を算出していると公表しているが，これらの特性値や尺度は具体的には公開されていない [15].

Intelligent Essay Assessor(IEA) は，1999 年，コロラド大学の Landauer らにより開発された [17]．これは LSA を取り入れている．語彙中に含まれる文字列ではなく，内容を重視している．百科事典や問題文に関係する専門書を用いたコーパス（Landauer らは”bag of words”と呼ぶ）を利用して，単語の共起から，内容を規定する特異値行列を与えることができるとしている. 内容，文体（首尾一貫性と文法），技巧（句法，スペル）という 3 つの観点に加え，総合スコアを A-D，F で示している．事前に人間が採点したデータを入力する必要があるが，数が少なくて済むという特徴がある．

e-rater は，世界最大の教育試験サービス期間である ETS（Educational Testing Service）が提供する AESS である [11]．アメリカの GMAT の小論文や TOEFL 試験などで広く利用されており，もっとも古くから長く使われている代表的なシステムである．開発当初は，構造／組織／内容に関する 60 変量を用いていたが，2004 年に公開されたバージョン 2 では表 2.1.2 の通り，12 の変量となっている． これらに係る重み付けを経験則によって表中の「重み付け列」のように定め，重回帰モデルによりスコアリングを行っている [18].

BETSY （Bayesian Essay Test Score sYstem）は 2002 年に，メリーランド大学の Rudner らによって，開発された [4]．多変量 Bernoulli モデルと multinomial モデルの 2 つのベイジアンモデルを用いて，4〜6 段階の評定に分類される．予め専門家によって採点されたエッセイをそれぞれ，適切/部分的に適切/不適切に分け，それぞれに分類するために特徴量を決めておく．次にその特徴量が各分類スコアに出現する確率の計算（多変量 Bernoulli モデル），および回答者のエッセイに含まれる確率を計算し（multinomial モデル）分類する.

IntelliMetric は Vantage Learning 社が開発した．ルール発見アルゴリズムに基づく知識工学的なアプローチで採点する [19]．予め採点が終わっている模範解答を学習し，各採点ポイントのデータから，人間の採点者の採点ルールの判断を推定する．5 つの「評価基準」一貫性，内容，構成，文章の複雑さ，アメリカ英語への適用を評価スコア観点として，各々1〜6 点で

表 2.1.2: e-raterV.2 の各変量と重み付け

No	変量	信頼性	重み付け	相関 [1]
1	総ワード数に対する文法エラーの割合	0.07	0.05	0.16
2	総ワード数に対する語の使用法についてのエラーの割合	0.16	0.02	0.2
3	総ワード数に対する手順のエラーの 割合	0.36	0.07	0.34
4	総ワード数に対するスタイルについてのエラーの割合	0.43	0.08	0.55
5	談話（discourse）ユニットの数	0.48	0.21	0.65
6	各 ユニットにおける平均のワード数	0.32	0.12	0.17
7	当該エッセイの 6 点法によるコサイン類似度が最大となるスコア点	0.38	0.04	0.44
8	最高点（通常 6 点）を得たエッセイとのコサイン類似度	0.24	0.07	0.5
9	単語の繰り返しの程度を示す指標 : 全ワード数（token）に対する異なったワード種類（word type）の 割合	0.47	0.08	0.32
10	Breland らの単語頻度指標に基づく語彙の難易度	0.11	0.03	0.22
11	平均の単語長さ,	0.25	0.03	0.32
12	単語の総数	0.56	0.2	0.78

[1] 2 名のヒューマンスコア平均との相関

評価し，全体評価も 6 点満点で評価する．いわゆる 6 つの段階に分類するものである．

Jess は，石岡らにより開発された日本語対応の AESS である [13]．e-rater の構造，組織，内容を踏襲し，修辞，論理構成，内容という評価観点を設けている． 3 つの観点の重み付けを 5:2:3 として 10 点満点でスコアリングする．ユーザはこの割合を変更可能である．プロの評価者の採点ではなく，社説などを書くプロのライターの記述から統計量（例えば修辞を示すメトリクスとして文の長さや語彙の多様性など）の理想的な分布を求める．それらと比較して，外れ値がある文書は減点してスコアリングする．現在ではひととおり開発を終え，実証研究を継続している段階で，センター試験での完全な実用化に至っていないが，設問や解答の膨大なデータを蓄積しつつ，システムの精度を上げる試みをしていると推測できる．またソフトウェアのパッケージまたは Web 上で提供しており，利用することができる．

Bookette は，California Testing Bureau（CTB）によって設計され，大規模試験や，クラス内での運用で利用されている [20]．自然言語処理（NLP），ニューラルネットワークを使用

しており，専門の評価者の手動採点結果をモデル化している．学生が作成した記述文の属性，例えば，文構造，単語選択/文法の使用法などを組み合わせて，効果的な文章の特徴を算出する [21]．このシステムでは，モデル化するために手動採点結果の多くのセットが必要とされる．また，文章レベルでの文法，スペル，ライティングに関するコメントを含むレベルなどの総括的なフィードバックをすることができる．

LightSIDE はカーネギーメロン大学で開発されたオープンソースのテキストマイニングツールである [22]．ユーザーが初心者であっても使いやすい GUI 環境を提供しているのが特徴で，AESS としての機能を有している．主に Weka（Hall et al.，2009）を実装し，教師あり学習により分類器を作成している．Naive Bayes と線形サポートベクターマシンの 2 つのアルゴリズムが用いられ，テキストの内容や文体などの複数の特徴を学習し，A・B・C・D などの成績ラベルに分類できる．Naive Bayes モデルでは，学習データに頻繁にみられる特徴が発生する確率を推定し，ラベル付けを行う．線形サポートベクターマシン分類器では，学習データから各ラベルのペアを総当たりで比較し，区別するための重みを計算する．これらの全重みを最終的に調整して分類器を作成している．

AI grading は，世界的に利用されているオンライン公開講座 MOOCs（Massive open online classrooms）で利用されている AESS である．MOOCs の第 1 世代では，数値，真/偽，および多肢選択式の回答のみであったが，第 2 世代からは評価方法が工夫され，自由回答方式の評価が可能になった [23]．具体的には，自己採点（self-assessment）・受講者の相互評価（peer assessment）・自動評価（AI grading）により，総合的に評価される．自動評価では，講師 (授業のインストラクター）が採点した中の 100 件のデータを訓練データとして選択し，機械学習を行う．当初は，スペル，文法，テーマのみであったが，近年ルーブリックを導入できるようになった．講師はルーブリックをテーマが異なる問題については，精度が低いことが実験で示されている．

KASS（Korean Automatic Scoring System for Short-answer Questions）は，韓国で開発されている短答式自由記述文の AESS である [24]．CSAT（大学修学能力試験）と NAEA（国立アセスメント）などの大規模試験における利用を目指している． また外国人に年 4 回提供される韓国語の試験 TOPIK（韓国語能力試験）も視野に入れているが，実用段階には至っていない．KASS は人間の採点者が作成した模範解答との類似度を評価するものである．情報検索技術（コサイン尺度，ベクトル空間モデル，潜在意味解析）を使用している. スコアリングでは，特定のスコアリングガイドラインで自動化されたスコアリングモデルの答えを適用している．問題のカテゴリーごとにスコアリング用のテンプレート（学生の模範解答）を利

用して配点する．このテンプレートはデータベース内に蓄積され，過去にない模範解答があれば更新される．Jess と同様に，短答式解答への対応を進めている [25].

AutoScore は，アメリカの非営利団体 American Institutes for Research によって開発された [20]．得点の高いエッセイと低いエッセイの訓練データが必要で，高低特性とを区別する概念を意味論に基づき学習する．例えば，段落内や段落間のつながりや一貫性，語彙や構文についても特性とする．また明確で命題ベースの迅速なルーブリックが利用可能な場合，システムは Proposition Scoring Engine に基づく測定値を統合することができるとしている．ただし，詳細は公開されていなない．

Lexile Writing Analyzer は，MetaMetrics によって開発された Lexile Writing フレームワークの一部である [20]． スコア，ジャンル，プロンプト，句読点に依存せず，意味論的複雑性（使用される単語のレベル）に関連する要因に基づいて，文章表現能力を予測するものである．このシステムでは，ライティング能力の近似値を表す少数の特徴量が使用される．各個人の基礎的なライティング能力を認識するため，エッセイ評価のための学習データは不要である．

GRASE は，Pacific Metrics 社が開発した．試行，属性抽出，スコアリングの 3 段階のスコアリングプロセスを実行する[5]．属性抽出のステップは，アイデア，文章の流暢性，組織，音声，言葉の選択，慣習，およびプレゼンテーションである．事前に採点された生徒のサンプルを分析して，スコアリングを行う．CRASE は，Web サービスとして動作する Java ベースのアプリケーションである．機械学習モデルを構築するために使用される構成と，人間と機械のスコアリング（すなわち，ハイブリッドモデルの導出）の融合としてカスタマイズ可能である．このシステムでは，エッセイを改善するために使用できるテキストベースおよび数値ベースのフィードバックも生成される．

2.1.3 日本語を対象とした AESS の研究

Jess ほど大規模ではないが，国内では 2000 年頃から多くの研究者が様々なアプローチで研究している．採点方法の理論的な提案や，各研究者の対象分野や組織内での限られたテスト試用となっており，汎用的な試験での実用化には至っていない．システムを開発する主旨についても，論作文指導，剽窃の検出（コピー&ペーストのチェックなど），採点支援（教員の採点負担軽減や評価の厳正化を目的として情報を提供するなど），ブレンド型の採点，完全自動採点を目指すなど，様々である．以下に関連研究を例示する．

論作文指導に主眼を置いた研究として，津森ら (2003)[26] の提出レポートの授受と管理，自

[5]https://github.com/target/grease

動採点を Web 上で行うシステムが挙げられる．学生は仮提出時に，章立て／キーワードの利用／文体に関する自動採点結果を受け，合格すると正式な提出を行う．これにより学生はレポートの形式を整え，記述事項を明確にする訓練が可能になる．教員側はその過程を確認でき，レポート作成指導に役立てることができる．しかし章立てのある決められた形式など，レポートが限定される．泉谷ら (2010) は，各採点項目を予測した結果を非採点者に提示するだけでなく，評価に対する根拠（説明文）も示すことで，支援を可能としている.「論点の言及」,「論理性」の評価に的を絞り，決定木によって支援情報を提示する [27].

剽窃の検出関連では，遠西ら (2008)[28] や渡邊 (2008)[29] の研究があげられる．前者は N-gram モデルを利用して提出者のレポートと Web 上の文章や他者が作成した文章との類似度解析を行う．後者は，機械学習を用いて，学生の良いレポートから教師データを作成し，新たに投入されたレポートを類似度により判定する．具体的には TF-IDF 法によってコピーレポートかどうかを判定し，理解度チェック単語数によって，よい考察か考察不足かを評価する．さらにこれらの値からニューラルネットワークの教師データを作成し，レポートを 5 段階評価するものである．教員評価との誤差が 15%とある程度の精度が確認できている．

採点支援目的では，椿本ら (2010) の研究が挙げられる [30]．LSA，クラスタリング，多次元尺度法を利用して，レポート内容の類似度を求めて多次元マップに可視化する．人間が採点する際提示することで，評価のばらつきを抑えることができる．

ブレンド型の採点では，村田ら (2008) の SVM（Support Vector Machine）を利用した小論文の採点支援システムが挙げられる [31]．人間によって採点された少数のサンプル答案とその採点結果を設問の採点項目別にそれぞれ解析し，得られた特徴量と採点結果のペアを SVM に学習させることで残りの大多数のテスト答案の評価値を予測する．さらに人間が GUI 上で各評価値の確認と編集を行うことにより，答案の最終評価を行う．設問の背景知識がなくても実用に足るだけの安定かつ高精度な自動採点が可能であることを確認しているとのことである．

最後に自動採点を目標とする研究では，採点のアルゴリズム，採点項目に関する着目点（形式，修辞，内容他）など，様々なアプローチで研究されている．藤田ら (2010) は，文章構造を解析し，小論文を論理性に関して採点する [32]．Jess に倣い接続表現を用いるだけでなく，主題連鎖関係，語彙連鎖関係，照応関係続関係に着目する．文間関係判定器には，SVM-Light を用いている．また，勝又ら (2013) は，SVM により，4 名の専門家が付与したスコアを基に学習データを作成し，論理構成の整然さについて自動評価する手法を提案している [33].

以上のように，多くの研究者が試行錯誤を繰り返している. 何れも精度を高める，すなわち

人間のスコアとの一致を目指すものである．利用している技術は，開発の目的により異なる．現段階では，限られた範囲内での試用にとどまっている．

2.1.4　採点手法

既存の AESS で利用されている採点手法は，何れも教師用データ（採点済みの記述文）から複数の特徴量を抽出し，それらをもとに，様々な手法で推測あるいは分類しながら評価値を判断するものである．表 2.1.1 に出現する「手法」は次ように整理することができる．

- 統計的手法による数値予測
  - 回帰分析：　e-rater に代表される．蓄積された採点結果から特徴量を抽出し，それらから重回帰モデルにより得点を求めるための回帰係数を求める．
- 分類による評価値の判定
  - ベイズアプローチ：　採点済みデータを幾つかに分類し，それぞれの分類を特徴づける特徴量を求め，それらが含まれる確率を事前確立としておく．それぞれの特徴量ごとに，採点するエッセイの事後確率を求める．
  - ルール発見：　採点済みのデータから，評価値を決める際の人間の採点ルールを調べ，推定値を決めておき，新たなデータを評価する際，それをもとに分類する．
  - 教師あり機械学習：　採点済みデータから，SVM（Support Vector Machine）などで学習して分類器を作成しておく．
  - LSI（潜在的意味解析）：　単語頻度や単語重要度から文書ベクトルを作成し，模範解答との類似度を計算して採点する．機械学習での過学習を抑えるねらいがある．
  - ニューラルネットワーク：　NLP（自然言語処理）により，単語だけではなく前後関係や文脈を対象に文をベクトル化して，正解をモデル化したものとの誤差により判定する

2.2　ルーブリックに関する先行研究

2.2.1　ルーブリックと AESS

ルーブリックは学習の到達度を測るための評価基準を表で示したものである．具体的な学習目標を示す観点と，その到達度を示す尺度および説明文からなるマトリックス型の表である．レポート，論文，プレゼンテーションなどのパフォーマンス評価のための最適なツール

と言われ [34]，多くの教員が導入している．ここでは採点の評定ガイドを意味する. 表 2.1.1 の「ルーブリック」欄で示したとおり，手動採点で利用するルーブリックをシステムに導入している AESS は e-rater と AI grading である．e-rater は専門の評価者 2 名が採点し，スコアに 2 点以上の差がある場合は，人間の評価者が調整する方式をとっている. この際，評価者はルーブリックを基に評価しているが，コンピュータシステム側の採点ではこのルーブリックを使用していない．MOOCs で利用されている AI grading は，自己評価，相互評価 (ピア評価)，AI 評価を組み合わせて総合評価を行っている [35]．授業担当者（インストラクター）がルーブリックを定義し，3 種類の評価で共通に利用することができる．授業担当者は評価の観点となるルーブリックを提示することができ，受講者はそれぞれの評価の観点について 5 段階評価とコメントを付す．ルーブリックに記述された文章と受講生の記述文の類似度を測ることができる．またルーブリック内の単語，あるいは表現にコストを付すことができ，より細かい採点が可能である．このシステムでは，課題ごとに内容が異なり，インストラクターが模範となる記述文を入力する必要がある．しかしながら，文献 [36] で例示されているルーブリックは，回答のポイントを文章で記述されたもので，評価観点と各グレードが書かれたマトリックス形式ではなく，模範解答に近い内容である．したがって AESS に利用できる汎用的なルーブリックではない．

近年ルーブリックに基づく評価が重要視されるにつれて，多くの高等教育機関で利用されている LMS，例えば Moodle や blackboard には教員が作成したルーブリックを入力し，評定結果を学生に提示する機能がある．また Google Classroom にも同様の機能があり採点効率や学生へのフィードバックを生かした研究が見られる [37]．しかしながら，これらの自動採点は選択式回答を中心としており，自由記述文の自動採点は現段階では見られない．

その他の AESS についても，機械の採点が人間と全く同じルーブリックを利用している，あるいは人間が利用するルーブリックに基づいて採点項目を設定しているものは見当らない．

2.2.2 ルーブリックの作成

ルーブリックの動向に関する文献 [38][39] によると，米国では全米カレッジ・大学協会（以下 AAC&U）がバリュープロジェクトを立ち上げ，機関を超えて活用可能なルーブリックを，1 年以上かけて作成している．これは探求と分析力，批判的思考力など基本学習成果として 15 領域をあげ，各々のルーブリックを作成したものである[6]．

日本では，学習指導要領改訂により初等中等教育でルーブリックによる成績評価が進んで

[6]https://www.aacu.org/value-rubrics

いるが，高等教育では現在も発展途上と言える [40]．大学など多くの研究機関で研修会が盛んに行われ [41]，成績評価への導入が推奨されている．日本高等教育開発協会では，教員個人が作成し，実際に授業で用いているルーブリックを自主的に提供し，共有するための場としてルーブリックバンクを Web 上に開設している[7]．

松下によるとルーブリックは，構造（分析的か一般的か），スコープ（課題を絞り込むか一般的か），スパン（長期か短期か）により，様々なタイプが存在する [42]．本研究では，初年次教育や基礎教育の半期授業で課される一般的なレポート評価での利用を想定しており，次の条件となる．

・構造：評価の観点と評価レベルをマトリックス型で複数設定する分析的な構造

・スコープ：学問分野や授業科目は特化しないが，課題がレポート方式に限られるという点で一部限定的

・スパン：短期間でスナップショット的に使われる採点用ルーブリック

以上の条件で，組織的なプロジェクトで作成・活用されているレポート評価に関する汎用性の高い既存のルーブリックとしては，関西国際大学のコモンルーブリック（ライティング)[8]，山口大学のコモンルーブリック [43]，および徳島大学の文章力ルーブック[9]などがあげられる．これらのうちの全文が公開されているものと，米国の AAC&U の文章コミュニケーション VALUE ルーブリック（Written Communication Value Rubric）[10][44](巻末の付録 図 A.3) の特徴を表 2.2.1 にまとめた．何れも，観点数は 4〜5，文章の体裁・文法・課題との対応・論理構成などを評価する．

2.3 語彙レベル辞書構築に関わる研究

本章では，語彙レベル辞書構築に必要なコーパスおよび単語の難易度算出に関する先行研究を紹介する．2.3.1 節で，コーパス構築の背景を，2.3.2 節で，文書の難易度測定に関する研究を，2.3.3 節で単語の難易度あるいは類似の指標を持つ語彙表に関する先行研究を紹介する．

2.3.1 コーパス構築の背景

コーパス構築の背景について，文献 [7] [10] [45] を参考にまとめる．コーパス構築は言語学の分野で始まり，最も代表的なものは，1961 年に構築された品詞などの文法的な素性を付

[7]https://www.jaedweb.org/blank-3
[8]http://renkei.kuins.ac.jp/pdf/3writing.pdf
[9]http://www.tokushima-u.ac.jp/cue/reform/ap/year/writing.html
[10]https://www.aacu.org/value-rubrics

表 2.2.1: 記述文評価のための汎用的なルーブリックの比較

ルーブリック（作成団体等）	観点数	評価観点内容	評価値等	備考
Written Communication Value Rubric（AAC&U）	5	文章作成の文脈と目的／内容の展開／ジャンルと学問分野の約束事／資料と根拠／構文と技法を操ること	0(1 に満たない場合) ／ 1(ベンチマーク) ／ 2，3(マイルストーン) ／ 4(キャップストーン)	ローカライズ（各大学や授業に合わせて変更）することを前提に，ある程度抽象化して作成
コモンルーブリック ライティング（関西国際大学）	5	課題に対する記述／論理的構成／レファレンス資料／文章の体裁／表現の推敲	レベル 0〜4	下位学年用，上位学年用と分け，達成レベルを変更．科目に応じて観点を追記
文章力ルーブリック（徳島大学）	4	主張の根拠付け／構成の明快さ／文章表現の適切さ／出典表示など	3 段階の尺度 (A) 結構です， (B) まずまずです， (C) 努力しましょう	レポート評価よりも，学習の振り返りを目的に作成

与したアメリカ英語の均衡コーパス Brown Corpus (約 100 万語) である．その後，イギリス英語の British National Corpus （BNC， 約 1 億語），同規模のアメリカ英語の American National Corpus（ANC）が構築された．1980 年代以降，辞書，新聞，書籍などの電子化が，さらに 1990 年代の Web 情報の増加に伴いスクレイピング技術が進み，多様な言語資源が利用できるようになった．現在，品詞以外に，統語構造や意味構造などの情報を付与した様々なコーパスが構築されている．日本では 1980 年代後半から，自然言語処理のためのコーパス構築が始まった [8]．1986 年，日本電子化辞書研究所（EDR）のプロジェクトにより機械翻訳を目的とした EDR コーパスが構築された．その後，新聞記事をもとに，形態素情報，統語構造，語義などの情報を付与したリアルワールド・コンピューティング（RWC）コーパスが構築された．1990 年代には，辞書や新聞の電子化テキストを用いて，形態素や統語構造，語義，照応などの情報を付与した京都大学テキストコーパスが構築された [46]．2011 年には，国立国語研究所を中心に「現代日本語書き言葉均衡コーパス」（BCCWJ）が構築された [47]．これらをもとに砂川らは日本語教育に必要と考える単語，全 17,920 語を抽出し，日本語教育語

彙表を構築し [48]，公開している [49]．また，文献 [50] では，BCCWJ を訓練データとして書籍やウェブページの分析を行い，コロケーション情報を生成・付与した難易度別コロケーション辞書を構築している．一方で，Wikipedia のような非均衡コーパスからシソーラス辞書を作成する研究もある [51]．このように，何らかのコーパスを利用して新たなコーパスを作成するなど，応用指向のコーパス構築が盛んとなってきている．佐藤は文献 [52] で，均衡コーパス BCCWJ の頻出語のカバー率を調べ，テキストの難易度セグメントによって，頻出語の集合が異なるため，元となるコーパスの選定には注意が必要であると述べている．

2.3.2 文書の難易度に関する研究

文書の難易度を測るには，文書全体の読みやすさを測る，テキスト（特に文字ベース）に着目して計算する，あるいは単語の難易度に着目する研究がある．日本語の自動採点システムとして先行している Jess では，採点基準に語彙水準を設定せず，類似の特徴量としてビッグ・ワード（big word，長くて難しい語）の割合を採用している [15]．これは，名詞の読み（カナ表記）の長さが 6 文字を超える割合から文書全体の読みやすさを測るものである．佐藤は，リーダビリティを表す公式に，単語ではなく文字を適用して求める難易度ツールを開発している [53]．

単語難易度に着目する先行研究として，自動採点システム e-rater ver.2[11] があげられる．英文の自由記述文の自動採点システムとして最も古くから開発され長く運用されているシステムの一つで，TOEFL などの大規模試験で運用されている．このシステムでは，単語頻度指数をもとに語彙レベル（a measure of lexical level）を算出し，採点特徴量の 1 つとしており，単語頻度指数が単語の難しさの指標となっている．これは，Breland[54] が，4 種のテキストコレクションコーパスを用いて単語頻度指数と単語難易度ランクとに高い相関関係があることを示した研究成果に基づく．

2.3.3 日本語語彙表と単語難易度に指標関する研究

単語難易度の定義や算出方法は複数提案されている．単語難易度を示す指標として，単語親密度，単語重要度，単語出現頻度，単語出現確率など様々な値や名称が用いられ，研究されている．これらの値そのものを難易度とする場合や，1 つの指標として利用し難易度を算出する場合がある．日本語教育の分野では，単語難易度を示す指標として単語親密度を用いている．アンケートから心理的尺度で親しみやすさを求める方法，単語テストや頻度から統計的に単語へのなじみを調べる方法などがある [55]．また，先述した BCCWJ や Wikipedia コーパス内の出現頻度をもとに各単語に難易度あるいはランク付け情報を付与して示すなど，

日本語の語彙表が構築され公開されている．これらの特徴を表 2.3.1 に示す．

表 2.3.1: 難易度指標を含む日本語の語彙表

語彙表（公開年）	単語数	難易度関連指標と算出方法
日本語教育語彙表（2012）[10]	17,920	『現代日本語書き言葉均衡コーパス』（BCCWJ）と「日本語教科書コーパス」より作成．難易度レベル 1〜6 を頻度などの指標をもとに設定し，人手で調整．
『現代日本語書き言葉均衡コーパス』短単位語彙表 Ver.1.0（2013）[11]	185,137	書籍，雑誌，新聞，ブログ，教科書など複数ジャンルのコーパスから単語を無作為に抽出．頻度の最も多いものからランク付け．
Simple PPDB : Japanese（2017）[12]	571,023	日本語の平易な言い換え辞書に難易度を追加したもの．日本語教育語彙表に由来する 3 段階（初級・中級・上級）の難易度を指定．日本語教育語彙表にないものは頻度に閾値を設け推定．

[10] http://jhlee.sakura.ne.jp/JEV.html
[11] https://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html
[12] https://github.com/tmunlp/simplejppdb/

まず，砂川らの日本語教育語彙表 [48] は，BCCWJ の教科書コーパスをはじめとする均衡コーパスを基に，複数ジャンルのテキストから一般的な日本語教育に必要な難易度を付与している．初出年や日本語教育での位置づけ，出現頻度などをもとに 6 段階に分け，様々な要因を勘案し人手により調整して構築されている [56]．次に，国立国語研究所による『現代日本語書き言葉均衡コーパス』語彙表は，BCCWJ 内の複数ジャンルのコーパスから単語を無作為に抽出して頻度順に並べ変え，通番をランクとして付与している．単語数は日本語教育語彙表の 10 倍である．

梶原らによる Simple PPDB: Japanese は，日本語学習者の読解支援を目的として，Wikipedia 内の出現頻度 5 以上の単語について，平易な言い換え辞書を構築したものである．日本語教育語彙表の難易度 6 レベルを初級，中級，上級の 3 段階に置き換え，新たに 1〜3 の難易度を付与している．日本語教育語彙表にないものは Wikipedia を分かち書きして単語を抽出し，頻度や単語長を特徴量として SVM による分類で難易度を推定している [57]．

そのほか，単語重要度を用いて，文書の専門性を評価する研究もみられる．滝川らは，専門的な分野に絞った特定分野のコーパスから単語自体の適切な重みを求め，専門辞書に単語重要度として付与している [58]．一般的に使われず，特定分野でも出現頻度が低い単語は重要度が高いとする．単語重要度の計算手法としては，単語の出現頻度をもとに計算する TF-IDF 値や，さらに TF-IDF 値を文書長で調整する Okapi BM25[59] が採用されている．作成した専門辞書を用いて文書（ツイート）の専門性スコアを計算し，ユーザーの専門性の高さを推定しランキングしている．

江原は，均衡コーパスや人手を介する言語資源に頼ることなく，生コーパスから直接的に難易度を推測する方法を提案している [60]．Wikipedia などのコーパスにトピックモデルの潜在ディリクレ配分法（LDA）を適用し，難易度指標を求める素性として，トピック内の単語の出現確率を用いる手法である．コーパスの単語頻度を素性として用いる従来法より，出現確率を用いる方が，単語難易度関連指標の予測精度が大幅に向上したことが報告されている．

第3章　ルーブリックを基盤とした評価モデル

本章では，本研究の特徴の一つであるルーブリックを基盤としたレポートの評価モデルについて説明する[1]．3.1 節で，ルーブリックを基盤とする理由を，3.2 節で，教員が手動で採点するレポート評価のためのルーブリックを提案する．3.3 節で，手動採点用ルーブリックをもとに作成した自動採点用ルーブリックを提案し，3.4 節で，自動採点用ルーブリックを基盤とした評価モデルについて説明する．3.5 節で評価モデルの妥当性を述べ，3.6 節でむすびとする．

3.1　ルーブリックの必要性

レポート評価では，採点者（評価者）による採点結果のばらつき，同一採点者内での採点の偏り，採点者の時間的負担など，様々な問題がある．多くの採点者はチェックリストや採点時の評価指標（いわゆるルーブリック）を定めて評価の厳正化を保つ努力をしている．したがって自動採点システムを構築するにあたり，こうしたルーブリックを基盤に採点のアルゴリズムを設計することで，手動採点と近い処理を実現し精度向上が期待できる．

また評価基準を教員と学生が共通に持つことが可能となり，学生にフィードバックできることから，学生・教員双方を支援するシステムとすることができる．石岡は，自動採点システムに望まれる要件の議論で，e-rater V.2.0 および Jess は論題によらず評価モデルは一定で，評価基準表に従った採点を行っている，として妥当性を述べている [63]．ここでいう評価基準表は，いわゆるルーブリックととらえることができるが，アメリカの経営大学院への入学試験である GMAT の採点基準を踏襲して設定している．

本研究では大学の授業で一般的に提出されるレポートの採点を対象としており，教育現場で実践的に活用できるシステムをめざしている．そこで，既存のレポート評価のルーブリックを綿密に分析・作成し，これに基づいて自動採点システムを設計すべきであると考える．

3.2　採点指標となるルーブリックの作成

表 3.2.1 にレポート評価のための手動採点用ルーブリックを提案する．先行研究の表 2.2.1 で示した汎用性のあるルーブリックを参考にした．なお AAC&U のバリュールーブリック (ライティングコミュニケーション）については，ウェブサイトで公開されている原文と松下ら

[1]本章は文献 [61] [62] を加筆・訂正したものである

の和訳 [42] を参考にした．

提案ルーブリックは 5 種の評価観点（Contents:課題の理解度と解答内容の妥当性，Structure: 論理的な展開，Evidence : 資料と根拠の妥当性，Style : 文章作法の遵守と適切な推敲，Skill : 読みやすさ・表現の巧みさ）を持つ．各評価観点の評価尺度として 5 つの達成レベルを設定し，レベル毎に 2 点の幅を持たせてある．尺度の説明文は，ルーブリックに関する論文やルーブリックバンクなど Web 上で公開されているレポート採点用ルーブリックで多く見られる項目や表現を参考にした．手動採点では，5 つの評価観点の各々について，尺度として記述されている内容に当てはまる位置を確認する．各レベルの中でどちらの点数になるか決め配点する．例えば，論理的な展開について B レベルであれば，4 点か 5 点かを決めることになる．観点ごとに，0-9 の範囲で評価値が決まることになる．採点結果は離散値になるが，3〜5 の段階評価となる他の多くのルーブリックに比べ，学生間の差が生じやすくなる．また各レベルに幅を持たせたのには，採点者のストレス軽減の目的による．一つの値に絞ると，それが相応しい得点かどうか迷うであろうが，2 点の幅があるとスコアリングしやすいと考える．また一つひとつの尺度を何度も読むのはストレスがかかる．到達目標（到達レベル 9）の尺度内容を理解し，0 から 9 で評価できれば，負担軽減が期待できる．

表 3.2.1: 手動採点のためのルーブリック

評価観点	達成ベルと配点				
	0-1	2-3	4-5	6-7	8-9
I.Content 課題の理解度と解答 (記述) 内容の妥当性	解答内容が，課題とは無関係である．	課題を理解し解答しているが，誤りがある．	課題を理解し解答しているが，記述が不足している．	課題を理解し的確な解答であるが，改善の余地がある．	的確な解答である．関連用語を適正に用いている．改善の必要はない．
II.Structure 論理的な展開	記述内容にまとまりがない．	理論の展開に矛盾がある．	順序立てて理論を展開しているが，改善すべき点が複数ある．	順序立てて理論を展開しているが，説得力がない．	順序立てて理論を展開している．意見・主張があり，説得力がある．
III.Evidence 資料と根拠（エビデンス）の妥当性	資料を全く参照していない．根拠を示していない．	資料を参照していないが，根拠を示そうとしている．	参照しようとしている資料は相応しくない，または信頼性がない．	信頼でき，関連性のある資料を参照しているが，引用・参照の仕方に誤りがある．	当該の学問分野にふさわしく，信頼でき関連性のある資料を，うまく使いこなしている．
IV.Style 文章作法の遵守と適切な推敲	複数にわたってルールを守っていない．文章が全く推敲されていない．	ルールを守っていない，誤字・脱字，文体の誤りなどが複数ある．	大よそのルールを守っているが，訂正すべき点が複数ある．	訂正すべき点はないが，改善の余地がある．	よく推敲している．全く誤りがない．
V.Skill 読みやすさ・表現の巧みさ	文章が読み辛い．明らかに文章スキルがない．	文章が長すぎるなど，複数の改善すべき点がある．	文章が概ねまとまっているが，改善すべき点がある．	文章が読みやすい．語彙が豊富である．	読み手に明確に意味を伝えることができ，読みやすい．語彙が豊富である

3.3　自動採点のためのルーブリックへ

表 3.2.1 で示したルーブリックの評価観点は大綱的であるため，細分化して評価項目を設定し，コンピュータによる自動採点のためのルーブリックを表 3.3.1 のとおり作成した. 評価項目についても評価観点の決定と同様に，大学・団体などの組織レベルで研究開発された汎用性のあるルーブリックを参考に作成した．評価項目のうち，現段階でコンピュータによる自動採点が困難であると判断した項目は,「採点」欄に「手動」と示した．評価観点のうち Content，Structure， Evidence（以下 CSE）は教員の価値判断に強く依存するため，ほとんどの評価項目について汎用的で精度が高い自動採点は困難である．また教育現場においては，担当教員が注力して読み，判断すべき項目でもある．他方，Style，Skill（以下 SS）の多くの項目は，計測可能な量的データであるため，自動採点の対象となり得る．文章の体裁や文法，論作文スキルにかかわる内容のため，機械的に正解・不正解を判断できる．教員の見落としを防ぎ，厳正な評価が可能であると判断する．また，CSE のうち，Content の「1) 論題と記述の合致度」および「2）主要な関連語の存在」を当面の間，自動採点可能項目として位置付ける．

表 3.3.1: 自動採点用ルーブリック評価項目

評価観点	自動採点用評価項目	採点
I.Content 課題の理解度と 解答（記述）内容の 妥当性	1）論題と記述の合致度 2）主要な関連語の存在	自動
	3）出題意図の理解度 4）内容の総合評価 5）学修内容の理解度	手動
Ⅱ.Structure 論理的な展開	6）論理性の水準 7）意見・主張の妥当性 8）事実と意見の区分け 9）説得力	手動
Ⅲ.Evidence 資料と根拠 （エビデンス）の 妥当性	10）参照資料の質水準 11）参照資料の関連性 12）論拠資料の妥当性 13）図表への説明付加 14）引用量の妥当性	手動
Ⅳ.Style 文章作法の遵守と 適切な推敲	15）文体の統一性 16）誤字・脱字の排除 17）構文の妥当性 18）主述関係の妥当性 19）句読点の妥当性 20）冗長さの排除 21）表記ゆれ・曖昧さの排除	自動
V.Skill 読みやすさ・ 表現の巧みさ	22）漢字の使用率 23）文長の妥当性 24）語彙の豊富さ 25）語彙の水準	自動

3.4　評価値の推計モデル

本節では，評価項目の自動採点結果から，観点毎の評価値の推計および総合評価を判定するモデルについて述べる．図 3.4.1 は採点方針を示したものである．

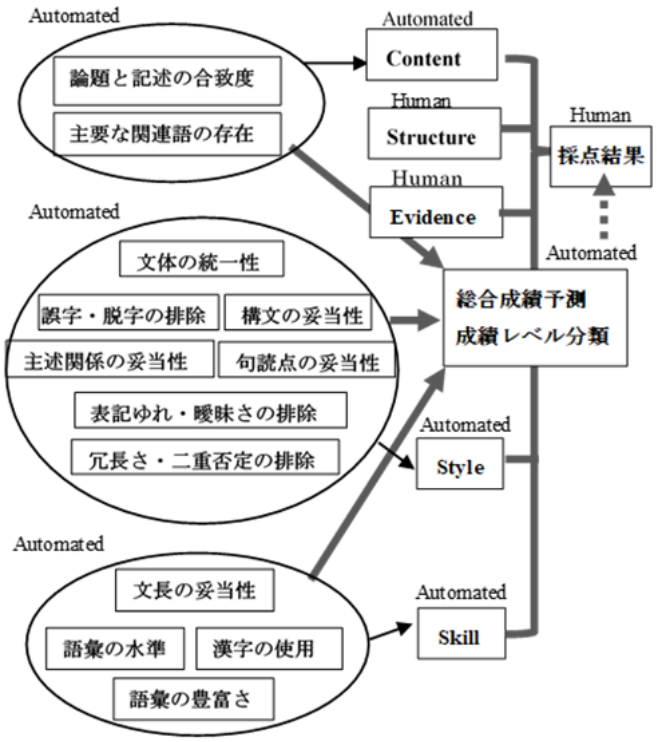

図 3.4.1: 総合評価の算出

全評価項目が自動採点可能である Style・Skill については，細分化した評価項目の採点結果を特徴ベクトルとする重回帰モデルにより算出する．回帰係数（各評価項目の重み）は採点済みレポートから算出する．

次に総合評価値を算出するモデルを述べる．教員が最終的に成績を判断する際，有用な資料として提示するために，総合成績レベル（A，A+など）を予測する．13 評価項目のうち「2) 主要な関連語の存在」については，教員が学生回答内に含むべきキーワードを設定しないケースもある．したがって推測値を求めるための特徴量からは省く．他の 12 評価項目の結果を基に，分類器によりレポートの総合点を推測する．具体的には，自動採点可能とするすべての評価項目の採点結果を特徴ベクトル (説明変数) としてサポートベクターマシン（以下 SVM）などにより，成績レベルを予測・分類する分類器を作成しておく．採点時にこの分類

器により分類した結果を採点者に提示する教員はレポートの内容や論理的展開など，CSE の自動採点されていない項目に集中して採点する．システムの提示する総合評価値を参考にしながら最終的な判断を行う．

3.5 評価モデルの妥当性

本節では，自動採点対象項目の採点結果から，他の評価観点を推測する手法の妥当性について確認する．表 3.5.1 は，教員 1 名が手動採点用ルーブリック（表 3.2.1）に従って採点した結果を，スピアマンの順位相関により相関関係を調べた結果である．レポートの氏名を伏せ，順序をランダムにして 3 回採点し，その平均を評価値（各評価観点のスコア）とした．

表 3.5.1: 手動採点での評価観点間の相関

	Content	Structure	Evidence	Style	Skill
Content	1	-	-	-	-
Structure	0.67***	1	-	-	-
Evidence	0.46***	0.59***	1	-	-
Style	0.54***	0.60***	0.47***	1	-
Skill	0.47***	0.62***	0.54***	0.72***	1

n=83, *** : $p < 0.001$

表に示す通り，CSE 間，SS 間，さらに CSE と SS 間で 5%水準で有意な相関がある. 下線は相関が高いことを示す．Skill についてはほとんどの評価観点で相関がみとめられた．Style 項目は CSE のうち Content との相関が最も高い. このことは，自動採点で求められる SS から CSE の評価点に関するある程度の情報，例えば総合成績レベルなどを推測できることを示す. 教員がレポートを採点する際，文章作成スキルに影響される，あるいは文章の読みやすさが内容を明確に伝えることに影響し，Content や Stracture の配点に影響を与えるものと考えられる．

3.6 むすび

レポート採点用の詳細な評価項目をもつルーブリックを提案し，自動採点が可能な評価項目と評価観点を明らかにした．特に Style 項目群のほとんどが，教員による採点結果との相関を認め，自動採点可能であることがわかった．また，5 つの評価観点間すべてについて相関を認めたことから，SS の自動採点結果により，CSE の予測が可能であることがわかった. ただし，Skill 項目群は，アルゴリズムの再検討もしくは，評価項目を追加して精度を高める必要がある． Skill 項目の精度向上の取り組みについては，第 7 章で議論する．

第4章　AES支援システムのアーキテクチャ

本章では，システムの全体像と，計算機内部における学生レポートを処理する過程，および前章で示した自動採点用ルーブリックの，各評価項目の具体的な計算方法について述べる[1]．4.1 節で，システムの全体像を，4.2 節で，自動採点評価項目の計算方法を述べる．4.3 節で，評価観点や総合評価の予測値の計算方法を述べる．4.4 節で，計算方法に関する議論を示し，4.5 節でむすびとする．

4.1　AES支援システムの全体像

図 4.1.1 に AES 支援システム全体像を示す．

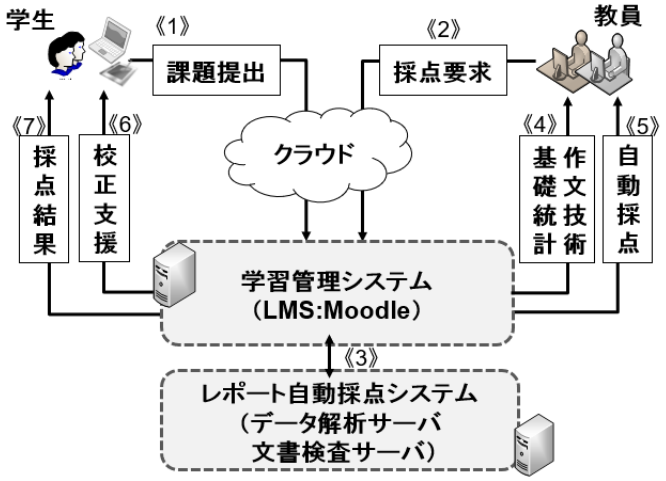

図 4.1.1: 自動採点支援システムの全体像

本システムは，学生のレポート作成能力育成および教員の採点を支援する LMS (Learning Management System) 上の AES 支援システムである．学生は Moodle の学習コースを通して課題レポートを提出する《1》．課題提出前に簡易的に誤字脱字や構文エラーなどをチェックすることができる《6》．教員は Moodle プラグインを操作し，採点対象となる課題モジュール名を選択して分析を開始する《2》．この際，クラス単位で処理を行うことができる．採点要

[1]本章は文献 [64] [65] [66] を加筆・訂正したものである

求が出されると，Moodle は外部サーバを呼び出してレポートの分析を行い，処理結果をユーザーに返す《4》《5》．また，採点結果を Moodle を通して示すことができる《7》．

システムの中核となる自動採点処理部《3》は，レポート（テキストデータ）の収集と学生へのフィードバックを担う LMS サーバ（Moodle），採点処理を担うデータ解析サーバ（TeMP），自由記述文の文章校正エラーチェックを担う文書検査サーバ（RedPen[2]）の 3 つのサーバで構成される．レポート自動採点プラグインは，梅村が開発した TeMP [1] の拡張モジュールとして実装した．Moodle に実装したプラグインブロックはユーザ処理のフロントエンドであり，自動採点処理部である TeMP および RedPen と接続する．RedPen は，伊藤が開発したオープンソースのドキュメント検査ツールである [67].

自動採点処理部の流れを，図 4.1.2 に示す.

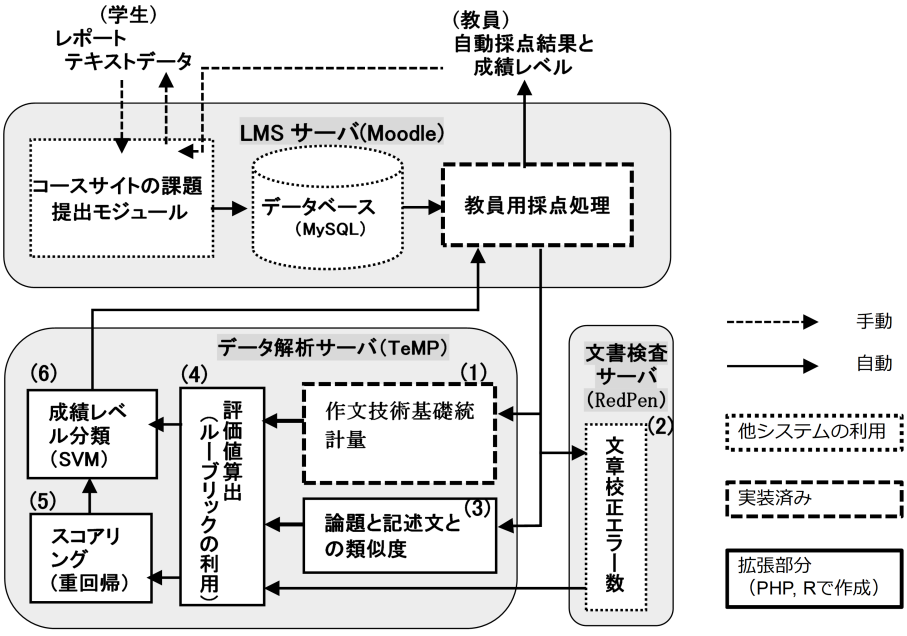

図 4.1.2: 自動採点システムの構成

[2]http://redpen.cc/index.html

図中の破線矢印は，手動の作業を示す．学生の作業として，レポートのテキストデータのアップロードおよび，提出前の文書チェック作業がある．教員の作業は，採点結果の要求である．あるいは，学生が直接アップロードしたデータではなく，外部ファイルをアップロードし採点結果を得ることも可能である．図中の破線枠は，すでに実装済みの TeMP が担う処理である. なお，プラグインの開発・実行環境（表 A.1），およびプラグインの実行画面（図 A.4) を末尾の付録に掲載する．採点処理プロセスに相当する (1)・(3)〜(6) は PHP と R によるプログラム群である．(2) の RedPen サーバとのやり取りは，REST API を利用する．採点モジュールに，重回帰モデル，コサイン尺度による類似度計算，サポートベクターマシンを用いる． システムは，(1) から (6) のプロセスを経て採点結果を提示する．図中の破線枠は各プロセスの処理内容は，次のとおりである．

(1) では，レポートを形態素解析し，索引語の出現頻度を示す文書行列の作成，文書ごとの文字数，文章数，トークン比など作文技術基礎統計量を算出する．形態素解析器は MeCab である．

(2) 文章検査サーバを用いて，文書ごと，評価項目ごとに，文体や文法上のエラー数を集計する．1 名分のレポートデータを RedPen に引き渡し，エラーの個数をプレーンテキストでまとめて受け取る処理を繰り返し行う．

(3) ベクトル空間モデルに基づき，教員の提示した論題と学生レポートとの類似度を，コサイン尺度によって計算する．

(4)(1)・(2) の結果から，ルーブリックに基づいた採点を行うために必要な評価値を，各採点項目の計算アルゴリズムにしたがって算出する．採点の基本的な考え方は，ルーブリックの適正基準との差異を，各文書の評価値とするものである．

(5) 学習データを用いて重回帰モデルであらかじめ求めておいた重み付けにより，(4) の評価値から Style と Skill の自動採点結果を計算する．

(6) 学習データを用いて事前に作成済みの分類器 (本研究ではサポートベクターマシン）により，(4) の評価値から，総合成績を A+・A・B・C・D の各レターグレードに分類する．

図 4.1.3 に (1) から (6) までの処理の流れを示す.

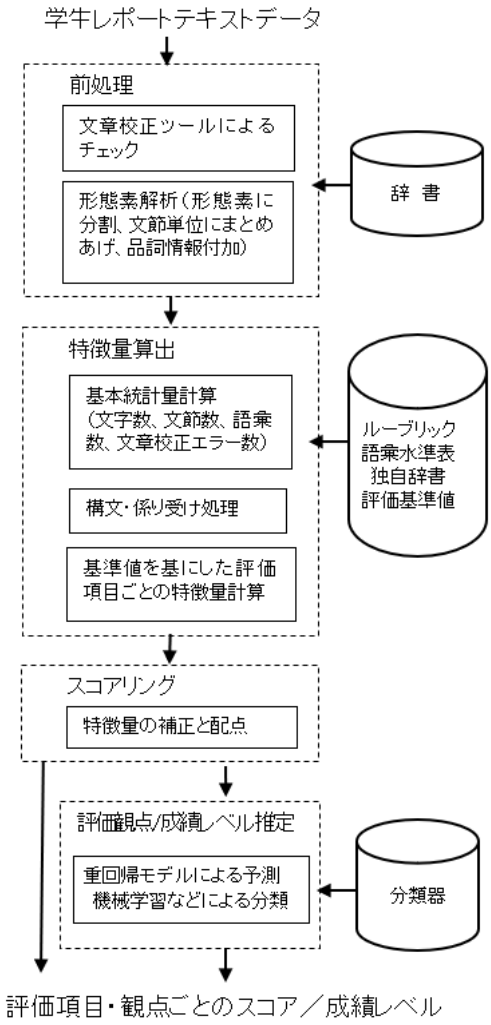

図 4.1.3: 自動採点の流れ

ここでの処理は，第 3 章の表 3.3.1 のルーブリックで提示した評価項目のうち，自動採点可能項目のみ，その評価値を算出することになる．LMS で入力された学生レポートのテキストデータは，教員がプラグインから採点処理を要求する (実行ボタンをクリックする) と，(1) と (2) の処理を並行して開始する．(1) は文章校正ツールにより基本的なチェックを行い，エラーの内容と数を算出する．また形態素解析器に付属の辞書を参照しながら，形態素に分類し，文節にまとめ上げ，品詞や文節位置など，評価項目の計算に必要な情報を付加する．

次に，ルーブリックで提示した評価項目ごとの計算アルゴリズムに必要な変数となる基本統計量を計算する．構文・係り受け処理では，構文の妥当性を計算するための変数を算出する．これらの変数から評価項目ごとの特徴量を算出する．特徴量とは評価項目のスコアリングのもととなる値である．例えば,「文体の統一性」であれば「非統一の割合」を意味する (表 4.2.1 の No.15 参照)．計算の過程で評価項目によって，語彙水準表や独自辞書を随時利用する．独自辞書については，大学生レベルの書き言葉言語コーパスを現在開発中である．算出された特徴量と，あらかじめ設定してある特徴量の基準値（望ましい値）との差を求め，特徴量とする．

スコアリングでは，特徴量の意味するところに即して値が大きいほど高いスコアになるように補正する．例えば，語彙の豊富さは特徴量が大きいほど配点が高くなるが，誤字・脱字では特徴量が大きいほど配点が低くなる．さらにクラス全体の平均，最大値，最小値をもとに補正し，9 点満点で配点して表示する．

評価観点／成績レベル推定では，教員の手動採点結果を目的変数，自動採点結果を説明変数として線形重回帰分析を行い，重回帰式から評価観点の予測値を求める．

また，過去のレポート採点結果をもとにサポートベクターマシンで分類器を作成しておき，総合成績レベルとして，5 段階のレターグレード（A+，A，B，C，D）への分類を行い表示する．

4.2　評価項目の計算方法と精度

本節では，自動採点の 12 項目の計算方法について述べる．4.3.1 項では計算方法と判断基準を，4.3.2 項では計算精度について確認する．

4.2.1　評価項目の計算

自動採点対象となる評価項目の評価内容と計算方法，および判断基準となる適正値を表 4.2.1 にまとめた．

表 4.2.1: 自動採点項目の評価内容

No.	評価項目	評価内容	計算式	適正基準
1	論題と記述の合致度	提示された課題との類似度	コサイン類似度	1 に近い程良い
15	文体の統一性	「である調」または「ですます調」の統一の可否	非統一の割合=非統一の数/文章数	0
16	誤字･脱字の排除	誤字・脱字の有無	誤字・脱字出現率=誤字・脱字の数/文章数	0
17	構文の妥当性	構文のねじれの可能性の有無	構文エラー率=（二重否定の数＋節が深すぎる数＋曖昧な名詞接続の数）/文章数	0
18	主述関係の妥当性	主述関係が妥当でない可能性がある文章	妥当性でない率=一文に二回以上接続助詞「が」が出現/文章数	0
19	句読点の妥当性	句読点間平均文字数 (適正範囲 13〜17 文字)	句読点間平均文字数=全文字数/句読点の数	11〜15
20	冗長さの排除	ことばの重複や繰り返しの有無	冗長さ出現率=（重複語の数＋語の繰り返しの数）/文章数	0
21	表記ゆれの排除	表記ゆれの有無	表記ゆれ出現率=表記ゆれ出現数/文章数	0
22	漢字の使用率	漢字の使用率	漢字の使用率=漢字の数／全文字数	32%以上
23	文長の妥当性	文の平均の長さ	文の平均文字数=全文字数／文章数	26〜41 文字を適正範囲とする
24	語彙の豊富さ	トークン比	トークン比=異なり語数／述べ語数	値が高いほど語彙が豊富であるとみなす
25	語彙の水準	主要語彙の平均水準	主要語彙の平均水準=（全名詞，形容詞，動詞の難易度*当該語彙の重み）の平均	値が高いほど水準が高とみなす

Content「1. 論題と記述の合致度」の自動採点は，図 4.1.2 のプロセス (3) で行う．教員が提示した論題とレポート内容のコサイン類似度を求める．プロセス (1) で論題とレポート記述文を，それぞれ形態素解析器 MeCab を用いて形態素に分解し，抽出した「名詞・形容詞・動詞・副詞」を索引語とする．次に文書ごとに索引語の出現頻度 (重みづけした) テーブルを作成し，文書ベクトルとして類似度を求める．文書間の類似度を求めるには様々な手法が検討されている．浅原（2016）は，内容一致と表現一致の 2 つの側面があるとして，文書間類似度として用いられる各計量の意味や特性を検討している [68]．ここでは記述内容の類似度を求めること，先述したプロセス (1) で文書ベクトルが求められていることから，情報検索技術のひとつとしてこれらをコサイン尺度を求める式（4.3.1)[69] にあてはめ，類似度を計算する．ベクトルの類似度を求めるベクトル成分値の計算にあたっては，索引語の出現頻度を TF-IDF 法によって重み付けした．

$$cos(\vec{d}, \vec{q}) = \frac{\sum_{i=0}^{n} d_i \cdot q_i}{\sqrt{\sum_{i=0}^{n} d_i^2} \cdot \sqrt{\sum_{i=0}^{n} q_i^2}} \tag{4.2.1}$$

ここで，$\vec{d}$ は学生の記述文から得た文書ベクトル，$\vec{q}$ は論題の文書ベクトルである．したがって分子は，学生文書と論題文書の双方に出現する索引語の，重みの積の総和である．分母は，学生文書の索引語の，重みの二乗の総和と，論題文書の索引語の，重みの二乗の総和の積を示す．求められたコサイン尺度が 1 に近いほど，学生文書と論題文書が類似しており，論題と記述の合致度を示すとする．

次に Style や Skill の計 11 項目の計算方法を説明する．評価項目 No.15，16，17，18，20，21 については，基本的にすべて同じ方針で採点する．文書検査サーバ RedPen より返されたそれぞれのエラー数の合計を文章数で割ることにより，エラー率を求め，値が低いほど文法的なエラーがない，すなわち文章作成スキルが高いとしてルーブリックの 0-9 に合わせた配点をする．エラー率 100%の場合は，ルーブリックの配点で 0（ゼロ）であり，エラーが 1 つもなければ 9 となる．エラーの個数を求めるためのパラメーターを表 4.2.2 に示す．RedPen サーバ内の定義ファイルに予め設定してある．定義ファイルに設定するパラメーター (RedPen では Validator と表記）と検査内容を表 4.2.2 に示す．各項目の計算方法について説明を付記する．

表 4.2.2: RedPen サーバによる文書検査内容

評価項目	RedPen の設定 (Validator)	検査内容 [1]
15 文体の統一	JapaneseStyle	不統一の個数
16 誤字・脱字の排除	InvalidExpression	不正な表現 (単語や句）の存在
	SuggestExpression	文の不正な表現の使用
	Okurigana	送り仮名の使い方が正しくない
17 構文の妥当性	SectionLevel	セクションレベルの深さ 5 以上
	DoubleNegative	文の二重否定
	JapaneseAmbiguousNoun-Conjunction	格助詞の「の」と 名詞の組み合わせが連続する
18 主述関係の妥当性	DoubledJoshi	文の不正な表現の使用
	DoubledConjunctiveParticleGa	一文に 2 回以上，接続助詞の「が」が出現する
20 冗長さの排除	DoubledWord，Doublejoshi	一文内で 2 回以上，同一の単語を使用
	SuccessiveWord	同一の単語を連続して使用
	SuccessiveSentence	同一の文を連続して使用
	DuplicatedSection	著しく類似する節が存在 (コサイン類似度)
21 表記ゆれの排除	KatakanaSpellCheck	文書内のカタカナの単語の表記ゆれ
	JapaneseNumberExpression	係数表現のスタイルが一貫していない
	SuggestExpression	言葉のゆれ
	JapaneseAnchorExpression	章・節の参照 (係数) が一貫していない

[1] RedPen 公式サイト http://redpen.cc/docs/latest/index_ja.html に掲載の RedPen 1.10 ドキュメントより抜粋し作成

「15 文体の統一性」は，RedPen サーバより返された，文体が統一されていない箇所の個数を集計し，レポートの文章数で除すことにより，誤りがある割合を求める．値が大きいほど文章に誤りがある（統一されていない）として,「15 文体の統一性」の評価点を低くする．

「16 誤字・脱字の排除」は，RedPen サーバより返された，誤りあるいは望ましくない表現（辞書にない単語や句，略語の利用など）の使用の総数を，文章数で除し，誤りがある割合を求める．

「17 構文の妥当性」では，文の二重否定や，一文に 2 回以上，格助詞の「の」と名詞の組み合わせが出現する場合に RedPen サーバより返されたエラーの総数を，文章数で除し，誤りがある割合を求める．

「18 主述関係の妥当性」は，一文に 2 回以上，接続助詞の「が」が出現する，あるいは同一の助詞が一文で 2 回以上利用されている場合，主述関係が妥当でない可能性があるとみなし，誤り率を求める．主語と述語が 1 対 1 であれば読みやすい文章であり，複文やねじれの可能性が低くなると言える．

「20 冗長さの排除」は，RedPen サーバより返された，一文内で 2 回以上同じ単語が繰り返されたり，著しく類似する節が存在する場合のエラーの総数を，文章数で除し，誤りがある割合を求める．

「21 表記ゆれ・曖昧さの排除」は，文書内の表記ゆれ，例えばサーバとサーバなどや，表や節の三焦経数が一貫していない場合に RedPen サーバより返されたエラーの総数を，文章数で除し，誤りがある割合を求める．

「19 句読点の妥当性」では，句読点間の平均文字数を求め，適正範囲との差異により評価する．適正範囲は，最小適正値 11～最大適正値 15 とする．最小適正値は教科書の句読点間平均文字数，最大適性値は，社説の句読点間平均文字数である．レポート記述文の句読点間平均文字数の値により，以下の処理を行う：

・範囲内であれば妥当として，妥当性=1.0

・範囲より少ない場合，妥当性=1-(0.1 ×（最小値 11-レポートの平均句読点間隔）)

-10.0 以下は 0.0

・範囲より多い場合，妥当性=1-(0.1 ×（レポートの平均句読点間隔-最大値 15）)

10.0 以上は 0.0

「22 漢字の使用率」は，適正値 32%以上とする．レポートの漢字の使用率を求め，以下の処理を行う：

・32%以上であれば妥当として，妥当性=1.0

・適正値に満たない場合，妥当性=レポートの漢字使用率× 0.8

この適正値は，ランダムに抜粋した Web 上の 3 件の雑誌記事の漢字使用率，および日本語校正ツールを開発している株式会社ジャストシステムが公開している「読みやすさの基準」[3]を参考に設定した．

「23 文長の妥当性」は，文長の適正範囲 26〜45 とする．この範囲外の文書については，上限または下限との差異により以下の処理を行う：

・教科書平均 26〜社説平均 41 の範囲内は妥当として 1 .0，妥当性=1.0

・範囲より少ない場合，妥当性=1-(0.05 × (最小値 26-レポートの平均文長))

・範囲より多い場合，妥当性=1-(0.05 × (レポートの平均文長-最大値 41))

・差が 20 文字以上の場合，妥当性=0.0

この適正範囲は，マイクロソフト社の文章校正ツールを実行した際，読みやすさの基準として表示される文長および，先述したジャストシステムの「読みやすさの基準」を参考に設定した．

「24 語彙の豊富さ」はトークン比（異なり語数と述べ語数の割合）で判断している．例えば同じ語句を繰り返し記述している場合は，値が低くなる．語彙の豊富さをどのように測るかは多様な研究があり，本研究で検討した内容について 4.6 節で述べる．

「25 語彙の水準」は，レポート全体で使用している単語がどのレベルの単語かを判断する．意味語と言われる単語を抜粋し，それらの単語のレベルの平均値を求めて評価する．具体的には，砂川ら (2012)[48] が研究成果として提供している日本語教育語彙表 [49] を用い[4]，レポートから抽出した形態素に対して,「1. 初級前半」〜「6. 上級後半」の 6 段階の評定値を割り当てる．レポートに用いられている語彙のうち名詞・動詞・形容詞・副詞の索引語文書行列を作成し，上記語彙表に存在する語彙すべてのレベルを割り当て，平均を求める．

以上をまとめると，15〜18，20，21 は，文法的な誤りや避けるべき構文誤りなどであり，文書検査サーバで漏れなく検出することになる．教員の手動採点時に，手直しするであろう表記である．ただし「17 構文の妥当性」については明かな文法誤りでだけで判定することは難しい．そこで本研究での判定基準を決定した経緯を次節 4.3.1 で触れておく．19，22，23 は，誤りではなく，基準を決め判断が必要な内容である．本研究では，適正値あるいは適正範囲を決めておき，範囲内であるかどうか，あるいは適正値との距離により評価値を判断する．24，25 の語彙については，様々な判断要素や基準が考えられる．24 については 4.5 節 4.3.2 で

[3]http://support.justsystems.com/faq/1032/app/servlet/qadoc?QID=022387(参照日 2017 年 12 月 31 日)

[4]本研究では，日本語学習辞書支援グループ （2015）「日本語教育語彙表 Ver 1.0」（http://jisho.jpn.org/） を利用した．

トークン比を判断要素とした経緯を述べる．25 については第 7 章で計算方法の改善を含め詳述する．

4.2.2 評価項目の計算精度の確認

ルーブリックは評価観点毎の評定表であり，評価項目ごとの評価基準は設定されていない．そこで，評価項目の各々の精度を確認するために，0-9 の範囲で各評価項目の採点を手作業で行った．一方で，No.15〜25 の Style・Skill 項目群の自動採点を行い，手動採点結果との相関を求めた．各評価項目は正規分布とならなかったため，スピアマンの順位相関で求めた．表 4.2.3 に示すとおり，Style のほとんどの項目について，ある程度の相関が認められる．ただし，相関係数が低い No.20・21 については，評価のアルゴリズムや項目の在り方を改善する必要がある．一方，Skill 項目群では No.22 のみ，教員評価との相関が認められた．表に示すとおり，Style のほとんどの項目について，ある程度の相関が認められる．ただし，相関係数が低い No.20・21 については，評価のアルゴリズムや項目の在り方を再検討する必要がある．一方，Skill 項目群では No.22 のみ，教員評価との相関が認められた．

表 4.2.3: 教員の採点と自動採点の相関

Style 項目群	相関係数	Skill 項目群	相関係数
15	0.329*	22	-0.322*
16	0.335*	23	0.017
17	0.275	24	0.121
18	0.248	25	-0.144
19	0.375*		.
20	-0.051		.
21	0.190		.
Style 観点予測値	0.602*	Skill 観点予測値	0.210*

表内の係数値はスピアマンの順位相関に基づく．
*は相関が認められることを示す．
n = 43　　*：p < 0.05,　**：p < 0.001.

また，教員の採点結果を目的変数，自動採点結果の項目群を説明変数として線形重回帰分析を行った．さらに，教員の採点結果（実測値）と重回帰式から算出した予測値との相関を求めたところ，Style 項目群については 0.602 と高い相関を認めたが，Skill 項目群については 0.21 と低い値であった．Style の各評価項目の精度を上げることで，Style 評価観点の予測モデルによる採点が可能であると言える．

4.3　評価観点および総合評価値の算出

評価観点 Style と Skill 評価値および総合成績レベル A+〜D を算出・提示し，教員の採点支援を行う. これらは，第 3 章 3.4 節で示した評価値の推計モデルにしたがい，前節で求めた自動採点結果をもとに算出する．4.4.1 では評価観点 Style と Skill 評価値を，4.4.2 では総合成績評価値の算出方法について説明する．

4.3.1　評価観点 Style・Skill の計算

評価観点の算出は，重回帰モデルにより推計する．採点済みの 83 件のレポートから，R の関数 lm() により各評価項目の回帰係数を表 4.3.1 の通り求めてある．前節で求めた評価項目算出結果に各係数をかけ合わせて評価観点 Style および Skill の値をそれぞれ求める.

表 4.3.1: 評価観点評価値推測のための重みづけ

評価観点	自動採点用評価項目	重みづけ
IV.Style 文章作法の遵守と 適切な推敲	15）文体の統一性	0.20
	16）誤字・脱字の排除	0.23
	17）構文の妥当性	0.35
	18）主述関係の妥当性	-0.09
	19）句読点の妥当性	0.37
	20）冗長さの排除	0.06
	21）表記ゆれ・曖昧さの排除	0.14
V.Skill 読みやすさ・ 表現の巧みさ	22）漢字の使用率	-0.14
	23）文長の妥当性	-0.12
	24）語彙の豊富さ	-0.04
	25）語彙の水準	0.04

4.3.2　総合評価の計算

総合成績レベルは分類問題である．採点済みレポートを教師データとして機械学習で分類器を作成する．本研究では，分類精度が高いといわれる SVM により，成績レベル（A+，A，B，C，D の 5 段階のレターグレード）に分類する分類器を作成しておく．分類器は R の kernlab パッケージの ksvm 関数を利用して作成した．予め作成した分類器（R のオブジェクトモジュール）をデータ解析サーバ内の処理 (6)（図 4.1.2）で呼び出し，分類する．分類器を作成する手法は複数あり，どの手法を選択すべきかは議論の余地がある．以降で，分類手法として SVM

のガウシアンカーネルに決定した経緯を述べる．

ローネンやヤングは，文献 [3] [70] で，テキスト分類に関する分類器について，信頼できる方法がないことを言及している．比較実験が同じ条件（環境，テストコレクション，文書分割方法など）である必要があり，同一研究者で一連の比較を行ったときのみ，信頼できると説明している．さらに，この点を勘案した上で，最良の分類器は次の結論になると述べるにとどめている．文献 [3] の記述を引用し，本研究に係る部分を以下にまとめる:

・多くの研究者の研究結果から，最良のものは SVM，AdaBoost，kNN，および回帰による手法である

・Rocchio とナイーブベイズは機械学習による分類器の中では性能が悪い

・ニューラルネットワークと決定木による分類は，SVM と同程度の良い結果を示すときもあれば，悪い時もある

以上を参考に，本研究では SVM と決定木を候補として，43 件の学生レポートをもとに機械学習による分類を試みた．SVM はカーネルトリックにより非線形の分類が可能であり，複数のカーネルが用意されている．また，決定木は良好な分類ができる可能性があり，影響を及ぼしている項目を視覚的に確認できる．これらの SVM（3 種類のカーネル）と決定木を用いて比較したところ，表 4.3.2 の結果が得られた．表 4.3.3(SVM のガウシアンカーネル）や，図 4.3.1（決定木による分類結果）はこれらの分類器でテストデータを分類し，視覚的に示したものである．特に決定木では，A や A+に分類されるものは 1 件もなく，分類精度が低い．これらの実験結果から，分類精度が最も高く，誤分類率が最も低いガウシアンカーネルを採用する．なお，この検証にあたっては Weka3.8.1 を用いて，細かいレベルでパラメータを変更しながら確認している．

表 4.3.2: 分類器の比較

Classifier		Correctly classified rate	
		Training data	Test data
SVM	Gaussian (rbfdot)	0.801	0.536
	Linear (vanilladot)	0.655	0.571
	Polynomial (polydot)	0.727	0.464
Decision tree		0.764	0.571

表 4.3.3: SVM 分類器による分類結果

評価値 (総合成績レベル)	教員採点結果	正解件数	不正解件数	不正解の状況
A	1	0	1	B に分類された
B	21	21	0	
C	16	3	13	B に分類された
D	5	1	4	B と C に分類された
合計 (件数の割合)	43(100.0%)	23(53.5%)	20(46.5%)	

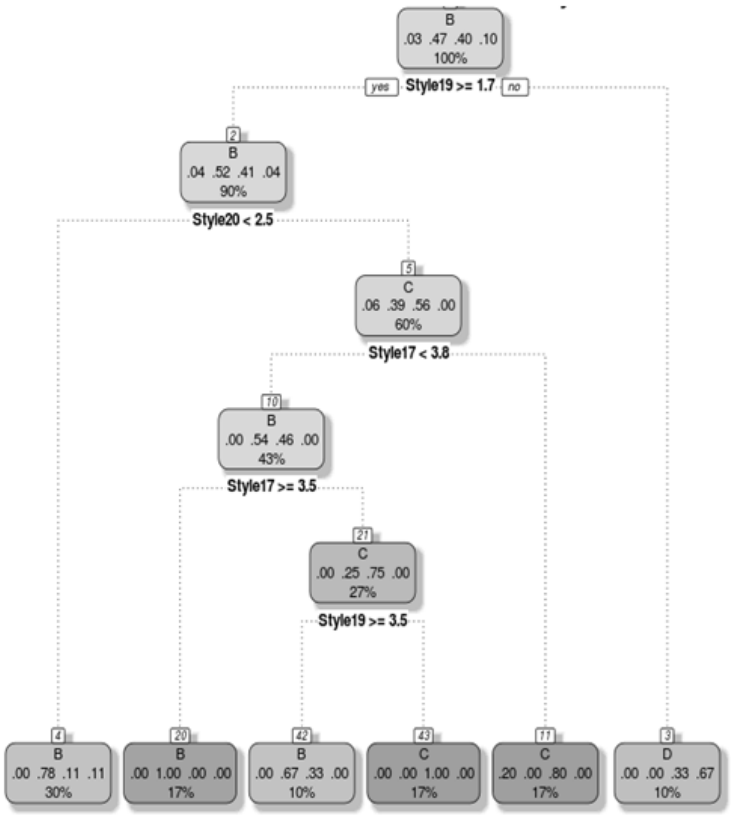

図 4.3.1: 決定木による分類結果

4.4　評価項目の計算方法に関する議論

「17 構文の妥当性」や「24 語彙の豊富さ」は，明らかな誤りから検出することはできない．したがって，評価の判断基準は複数考えられる．ここでは計算方法の決定に至る経緯を説明する．

4.4.1　構文解析の検討と課題

構文の妥当性を評価する方法として，次の 2 つのアプローチから検討した：

1) 文章全体の構造から，読みやすさを測る

二重否定となっている文章や，一文に 2 回以上，格助詞の「の」と名詞の組み合わせが出現する場合は，文章が読みづらくなる．例えば，前者の例は「アジア圏の旅行者は，買い物をしないことはない」，また後者の例は「アジア圏の旅行者の買い物の」などがある．何れも文章を読みにくくしている．

2) 文章の部分的な係り受けに着目し，読みやすさを測る

構文解析ツール CaboCha が出力する係り受け情報 [71] から，係り受け関係にある文節間の距離を算出する．この距離は，図 4.4.1 で示すように係り受け先が，いくつ先の単語であるかを示す数値である．

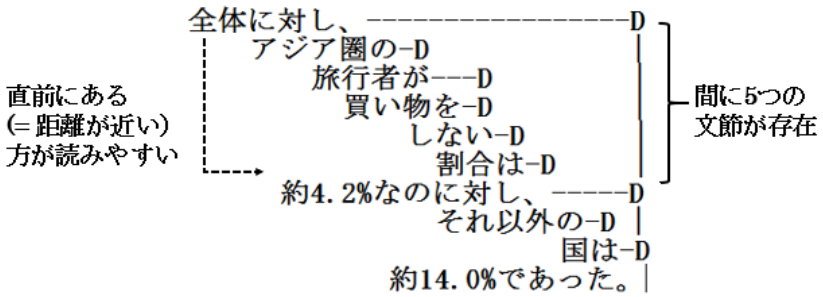

図 4.4.1: 係り受けによる構文の妥当性の採点（CaboCha の構文解析結果表示例）

次式のように，レポート内の数値の和を全文節数で除することで，各レポートの平均係り受け距離を算出する．

係り受けの平均距離 ＝ 1 文書 (レポート）の全係り受け数 ／ 文節総数

図 4.4.1 の学生レポートの一文を例にとり説明する．文頭の「全体に対し」の係り先は，6 つ先（距離=6）の「約 4.2%なのに対し,」である．本来なら係り先の直前に置き (距離=1)，

文章全体を「アジア圏の旅行者が買い物をしない割合は，全体に対して約 4.2%なのに対し，それ以外の国は約 14.0%であった.」とする方が読みやすい文章となる．例文では文節総数 10 であり，文節「全体に対し」の距離が 6，文節「約 4.2%なのに対し,」の距離が 2，文末は 0，それ以外は 1 である．したがって，この 1 文のみから成る文書の係り受けの平均距離は，1.5 となる．この距離の妥当性については定まった基準が見当たらないため，ランダムに検索した新聞記事 (2900 文字) から求めた平均距離 1.988 を適正基準と定め，この基準値からの偏差が大きい程，読みづらくなるとして低く採点する．

以上 1)，2) 各々について計算した結果と教員採点との相関を確認したところ， 1) の文章全体から見る方が，良い精度が得られた．そこで本研究では 1) を採用する．

4.4.2 語彙の豊富さの評価方法

Style・Skill 項目のうち,「24）語彙の豊富さ」の判断基準については，様々な報告がある [72]．本研究では文書の総語数・異語数・トークン比・ユールの K 特性値（以下，K 特性値）を指標として選定し比較検討した[5]．各指標の原理や計算方法は次のとおりである．

・ 総語数 (延べ語数)：文中に出現した単語の延べ数 (頻度の総和)．

・ 異語数（異なり語数)：文中に出現した単語の異なり数（同じ単語を 1 語と数える)．

・ トークン比：異語数／総語数によって求める．文書サイズ（文字数）に依存しない．

・ ユールの K 特性値 [74]：単語の頻度スペクトルと総語数を考慮し算出する．値が小さいほど,「語彙が豊富である」とみなす.

特性の異なる 2 種類の日本語ドキュメントをレポートとみなし採点した．一方は，毎授業時に実施する形成的授業評価記述文（レポート 1)，他方は面接模擬試験実施後の所感（レポート 2）である．前者においては，150 文字を目安に作成するよう学生に求めており，付与した作成時間も 5 分程度であった．後者は,「自身の模擬面接試験の録画を視聴し，800 文字相当の振り返りを準備して入力する」という設定のもとで作成させた．

当該コースを担当しない教員 1 名が採点にあたり，事前に作成した次の 10 項目からなる簡易的なルーブリックにもとづき，各項目を 0 点・0.5 点・1 点の 3 段階で評定した．

1) 文章が簡潔で読みやすい

2) 文体が統一されている

[5]本内容は，文献 [73] を加筆修正したものである．

3) 稚拙ではなく大学生らしい表現である

4) 文章表現が適切である

5) 誤字・脱字・言葉の誤りがない

6) 提示されたテーマに則している

7) 主張が明確である

8) 論拠が明確である（説得力がある）

9) 一貫性・整合性を保ちつつ論を展開している

10) 記述量が十分である

そして最終評定値として，これらの総和（上限 10 点）を求め，さらに「語彙の豊富さ」を表す 4 つの指標を算出した．表 4.5.1 は 2 つのレポートの特性および語彙の豊富さに関する各指標の平均・標準偏差を示す．

表 4.4.1: 学生レポートの基礎統計量

統計量	レポート 1（n=42）		レポート 2（n=25）	
	平均	標準偏差	平均	標準偏差
評定値	7.85	1.37	8.04	1.41
文字数	186.87	47.29	889.44	182.20
異語数	34.02	7.93	106.40	21.10
総語数	40.98	10.05	191.56	40.40
トークン比	0.84	0.07	0.56	0.06
K 特性値	118.53	76.28	147.32	40.71

評定値の分布を確認したところ，レポート 1・2 ともに正規分布に従わなかった．そこで，関連性を表す統計指標としてスピアマンの順位相関係数を用いた．分析結果を表 4.4.2 に示す．レポート 1 において，異語数・総語数に 5%水準で有意な相関を観察した以外は，評定値と指標間に関連性を見いだせなかった．2 つの指標について低度の相関を得た理由は，文字数が少ないために語数の多い文書が採点者の目に留りやすく，これが好印象となって評定値に影響したためであろう．なお石岡ら [75] は，K 特性値と得点率（試験の正解率）が無相関であったことを報告している. 本研究においてもこの点を確認できた．

表 4.4.2: レポート評定値と語彙指標との相関

	レポート 1 評定値（n=42）	レポート 2 評定値（n=25）
異語数	0.32*	0.10
総語数	0.42*	-0.02
トークン比	-0.26	0.05
K 特性値	0.18	-0.21

※ 表内の数値は，スピアマンの順位相関係数を表す．　　*: $p < 0.05$

4.5 むすび

本章では，前章で提案したルーブリックに基づく自動採点支援システムのアーキテクチャを説明し，システムの概要と処理の流れ，および評価観点や評価項目の具体的な計算方法について述べた．本システムでは，25 評価項目を手動と自動に区分する．自動採点結果を特徴量として，重回帰モデルで Style・Skill 評価観点の評価値を推測する．また，自動採点可能な 12 評価項目から機械学習により分類器を作成しておき，総合成績レベルを分類・提示することで教員の採点支援を行う．

第5章　AES支援システムの評価実験

本章では，自動採点支援システムを用いた実験結果を報告する[1]．5.1節で，実験の目的を，5.2節で実験で用いたレポートの概要を述べる．5.3節で評価項目および総合成績の採点結果を述べ，5.4節でむすびとする．

5.1　はじめに

ここでの評価実験は，システムの基盤となる自動採点部分の妥当性を確認することをねらいとする．学生が提出した小レポートを用いて，提案ルーブリックの評価項目のうちの自動採点可能項目を計算し，教員の手動採点結果と比較することで採点精度を確認する．

5.2　分析対象レポート

本システムで扱うレポートは，初年次教育や教養教育，リテラシー教育など，大学の基礎教育の授業で課す記述文や，エッセイタイプの100〜4000字程度の日本語論作文である．ここでは，表計算ソフトを使ったデータ解析をテーマとする情報リテラシー科目を受講する，大学1年次生が提出した小レポートを評価実験に用いた．レポートのテーマは,「外国人旅行客に関する調査報告書」の作成である．国土交通省観光庁のオープンデータをもとに表とグラフを作成し，そこからわかることなど自分の見解を，200文字以上で記述するものである．表5.2.1に，採点するレポートの特徴を示す．レポート全体での平均文字数は378.4，標準偏差201.5であった．

表 5.2.1: 採点対象レポートの特徴

クラス	人数（レポート数）	平均文字数（標準偏差）
A	43	427.6 (216.0)
B	40	325.5 (171.7)
全体の集計	83	378.4 (201.5)

[1]本章は文献 [66] [76] [77] を加筆・訂正したものである

5.3 自動採点項目の評価実験

教員はルーブリックにしたがい，手作業で評価観点の採点を行った．一方で，構築した AES 支援システムにより，評価項目 No.1 および，No.15〜25 の Style・Skill 項目群の自動採点を行った．自動採点の結果から，Style および Skill の評価観点の評価値を予測し，さらに，分類器により総合成績レベルに分類した．以下，5.3.1 項で，評価観点の採点結果を，また，5.3.2 項で総合成績レベルの分類結果について報告する．

5.3.1 評価観点の採点結果

評価観点 Style・Skill の教員採点と自動採点によるスピアマンの順位相関を求めた結果を，表 5.3.1 に示す．特に Skill 観点の値が小さく改善すべき課題であることが確認できた．

表 5.3.1: 重回帰モデルによる評価観点のクラス別予測結果の精度

クラス	文書数	平均文字数	手動採点との相関	
			Style	Skill
A	43	427.6	0.602	0.210
B	40	325.5	0.463	0.089
合計	83	378.4	0.493	0.255

5.3.2 総合成績レベル分類結果

分類器は今回の実験データで再作成した．4.4.2 項で示した方針で，SVM と決定木で確認した．何れも交差検定により行うとして，レポートの 2/3 に相当する 55 件をランダムに選択し学習データとし，残りの 28 件をテストデータとした．図 5.3.1 は SVM による評価結果と，分類器作成時のパラメータ及び精度である．また図 5.3.2 は決定木による分類結果である．やはり SVM の方が精度が高いため，SVM で分類器を作成した．

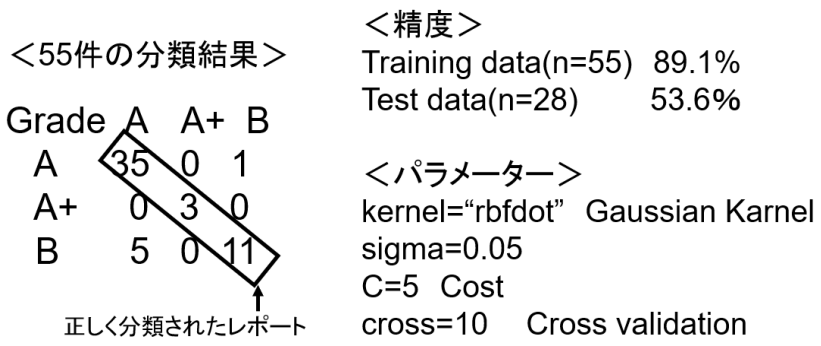

図 5.3.1: SVM による分類

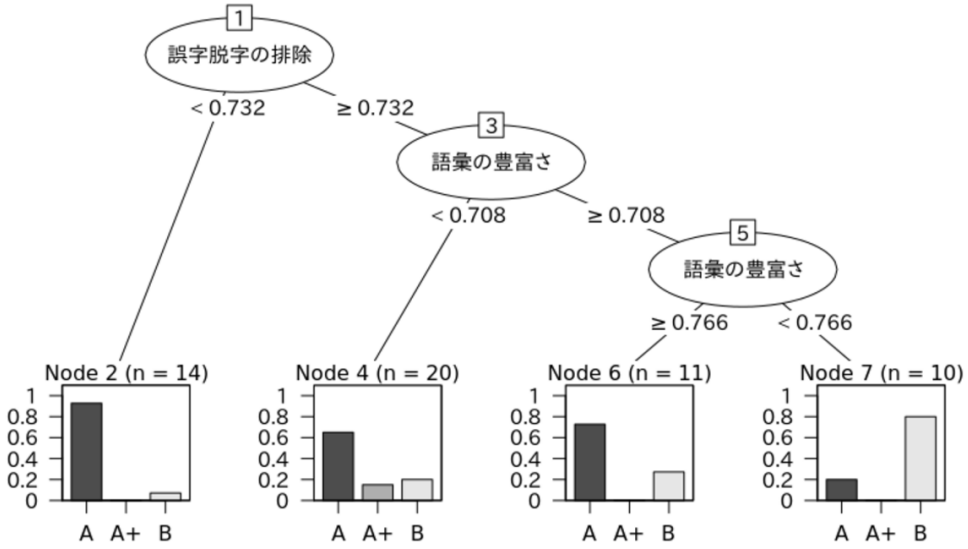

図 5.3.2: 決定木による分類

表 5.3.2 はクラス別の成績レベル分類結果の精度を示したものである．

表 5.3.2: クラス別総合席積レベル分類精度

クラス	人数（レポート数）	分類精度
A	43	0.581
B	40	0.550
全体の集計	83	0.536

図 5.3.3 は，教員が採点処理を行うと表示される画面のキャプチャである．Style・Skill の各評価項目から重回帰モデルで求めた評価観点 Style・Skill の各スコアを，ルーブリックと同様に，9 点満点に正規化して提示する．また成績レベルは，前節で述べた分類器により求める．

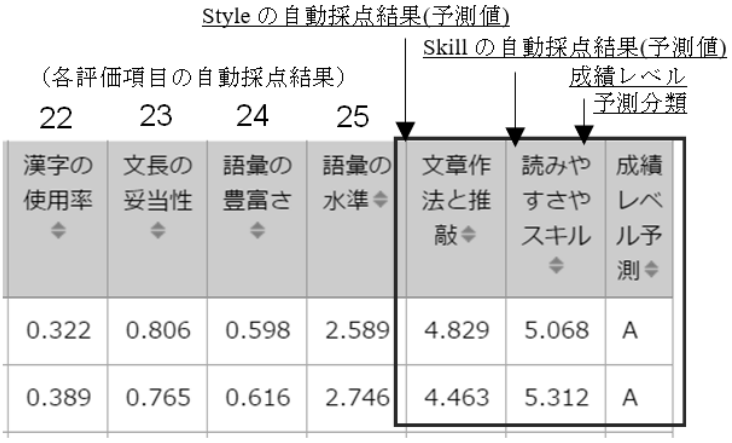

図 5.3.3: 採点結果アウトプットの例

5.4 むすび

レポート採点用の詳細な評価項目をもつルーブリックを提案し，自動採点が可能な評価項目と評価観点を明らかにした．特に Style 項目群のほとんどが，教員による採点結果との相関を認め，自動採点可能であることがわかった．また，5 つの評価観点間すべてについて相関を認めたことから，SS の自動採点結果により，CSE の予測が可能であることがわかった．ただし，Skill 項目群は，アルゴリズムの再検討もしくは，評価項目を追加して精度を高める必要がある．今後は SS の自動採点の精度を高める必要がある．提案手法では，評価項目を手動と

自動に区分して採点し，自動採点結果から手動採点部分を推測して成績レベルを提示することで教員の採点支援を目指している．またテキストデータの統計基礎値から，論作文指導に有用な多くの情報を得ることができる．今後は，採点結果や統計情報をより明確に提示するインターフェースの作成，テストデータによる処理速度の評価，新たな採点評価項目の検討，そして機械学習アルゴリズムによる成績分類精度の向上などを目指す．また，教育方針や授業方法に応じ，動的にルーブリックを切り替える技術の導入を検討する

第6章　教育現場での利用

本章では，AES 支援システムを教育現場で試用した結果について報告する[1]．6.1 節で本章の目的を，6.2 節でシステムの概要とシステムを利用するねらいを述べる．6.3 節で，情報リテラシー科目における教員の取り組み事例を，6.4 節で，学生のレポート作成時の利用事例を通して，AES 支援システムの有用性を確認する．6.5 節で，むすびとする．

6.1　はじめに

本節では，教育現場でのシステムの利用を通して，教員および学生それぞれの利用者からみた効果について考察する．レポートなどの自由記述文は，学習者の知識や，発散的思考力・問題解決能力等の評価情報を提供しうる点で有用である．一方でレポート作成指導や採点の厳正化が課題である．構築したルーブリックに基づく AES 支援システムで授業の課題レポートの採点をすることで，教育改善に結びつけることができる．システムの利用は，採点の正確性を高め，教員の時間的負担の軽減および，学生の記述文の傾向や授業の進め方など，授業改善に役立つ情報を得ることができる．

6.2　AES 支援システムの概要と利用のねらい

昨今の大学教育では，アクティブ・ラーニングの必要性や評価の厳正化が議論されている．このような授業での学修成果を従来の選択・穴埋め形式で評価することは困難であり，レポートやプレゼンテーション等の記述形式によるパフォーマンス評価が採用される．レポートに代表される自由記述文は，学習者の知識，思考力や問題解決能力等の習熟度を測ることができ，有用な評価方法とされている．レポート評価では，複数教員で一つの授業を担当する場合や，一教員が多数の受講生のレポートを採点する際に，評価のバラツキが生じやすい．そのため，チェックリストや評定ガイドを作成するなどして，採点の厳正化のために多くの労力を費やすことになる．またレポート作成指導の機会が増えると，学生と向き合うために必要な時間を確保するため，評価時間の負担減が課題となる．さらに学生自身がレポート作成スキルの習得に主体的に取り組むしくみも求められる．

本研究では，ルーブリックに基づく LMS 上の自動採点システムを提案し [62]，その構築を

[1]本章は, 文献 [78] を加筆・訂正したものである．

行ってきた [64]．現段階の開発システムは，作文スキルに関わる部分を自動採点し，内容に関する部分を手動採点するハイブリッド型システムである．システム導入により教員は，作文スキルの採点の正確性を高め，内容の読み込みに集中し，採点時間を短縮できる．また学生の記述文の傾向や，授業改善に役立つ情報を得ることができる．学生は，レポートの形式的なチェックや採点結果のフィードバックにより，主体的に学びながらレポート作成スキルを習得することができる．システムの中核であるルーブリックは，学生の学習到達度の判定だけではなく，評価の根拠や改善すべき点を示す教育ツールである．したがって，本システムを最終的な成績評価として採点に利用するだけでなく，授業内外で活用することで，様々な教育上の効果が期待できる．

本システムは第 4 章で示した通り，Web 上のアプリケーションである．梅村が中心となり開発した Moodle の簡易型テキストマイニング・プラグイン TeMP[1] を拡張して開発している．LMS（Moodle）は全学的に利用されているため，課題の収集やフィードバックの閲覧など，教員・学生ともに特別に操作上のストレスは生じない．学生はテキスト形式で Moodle にレポートの自由記述文を入力する．教員は Moodle のプラグインから処理したい記述文が入ったデータをクラス単位で指定し採点処理を依頼（実行）すると，Moodle からデータ解析サーバや文書検査サーバにデータが引き渡され，採点処理を実行する．結果と作文技術（レポート作成スキル）に関する各種統計情報が教員に返される．自動採点機能を追加することで，採点結果だけではなく，学生の語彙力や誤りの傾向などの情報を得ることができるしくみとなっている．特にクラス単位で処理が可能であることから，クラス毎の傾向を知ることができ，以降の学生指導や，授業運営に役立つ情報を分析しまとめることができる．

システムの特長は，レポートを評価するにあたり，表 3.3.1 に示すレポート採点用ルーブリックを基盤として，採点アルゴリズムを定めている点である．ルーブリックは採点のためだけではなく，学生に改善点を示す教育ツールとして役立てることができる．本システムは，採点結果だけでなく，改善に向けてのフィードバック情報を提供する．

6.3　AES 支援システムによる教育改善の取り組み　ー教員の利用事例ー

6.3.1　実践した授業概要と科目の位置づけ

本システムは，一般教養科目やリテラシー系などの基礎教育科目のレポートを対象としており，専門用語が多用される授業課題を対象としない．ここでは，2016 年度に担当した情報基礎教育の授業で提出されたレポートを，構築したシステムで採点し，出力された情報を分析するなど，教育改善につながる効果的な利用を目指した．

取り組みを行った授業は，表 6.3.1 授業概要 に示す『現代情報処理 B』（1 年次 2 期開講 2 単位）である．現代国際学部では 1 年次の情報基礎科目として『現代情報処理 A』（Word，Web ブラウザ，Power Point を中心とした PC 基礎とドキュメント処理）および『現代情報処理 B』（Excel による情報の整理・分析）を開講している ．情報基礎科目群全体を通して，大学での学びや卒業後の業務に ICT をツールとして活用できる能力を身につけることが狙いである．表やグラフはレポートで論拠を示すのに利用されることから，1 年次の基礎教育として重要な役割を担う．

表 6.3.1: 授業概要

科目名とテーマ	『現代情報処理 B』: スプレッドシートを用いた「データ分析基礎スキル」の習得
開講期，単位など	1 年次 2 期，演習 2 単位
学習目標	「データ収集・整理・分析」のプロセスを通じて次に掲げる学習目標の達成を目指す： (1) スプレッドシートの基本操作能力を習得する (2) 習得した技術を実データに適用しながら問題解決力を身につける (3) 収集したデータから得た知見をレポートに集約できる
目標達成度の評価方法	第 11 回・第 15 回に表計算ソフトのスキル確認問題とレポート提出
取り組みを行ったクラス（仮称）と受講者数	A（43 名），B（40 名），C（42 名） ※レポート提出者の人数．A・B クラスは 2015 年度，C クラスは 2016 年度開講

6.3.2 授業運営上の問題点と改善内容

ここ数年当該授業を担当することにより，レポートを書くことが苦手，あるいは気が進まない学生が多いことがわかった．また外国語大学に所属する学生の多くは外国語教育に興味がある一方で，情報教育への関心が薄い，あるいは PC に対する苦手意識が高く，モチベーションが低い．さらに大学あるいは学部全体の共通科目としての位置づけから，他の専門科目のレポート作成やゼミナールでの研究活動との連携が困難で，情報基礎スキルの習得に留まりがちである．しかしながら本科目は，大学 1 年次に習得すべきアカデミックスキルとして重要であることから，理解度やモチベーションを高める必要がある．これまで授業改善の

ための工夫として，LMS（具体的にはMoodle）を学生とのコミュニケーションの基盤として位置づけ，教材や授業スライドの事前提示，授業評価アンケート，授業内作業ファイルの提出などを行ってきた．また随時授業所感の収集を行い，学生の理解の状況および授業の進行度やモチベーションなどを把握することで，対面コミュニケーションに代わる一人一人の状況把握を行ってきた．教材としては，政府が公開しているオープンデータを利用し，観光局のデータによる動機づけ，および自らが知識発見をできるような課題づくりに取り組んできた．

しかしながら受講人数が多く，採点の正確性を高めるためにチェックリストを作成するなど時間的負担が大きく，授業実施期間内に学生の目標到達度の確認や，作文スキル改善に向けた情報をまとめ，フィードバックするような十分な時間が得られなかった．また授業所感を見るとクラスによる違いが漠然とわかるものの，具体的にどのような指導が必要かは，明確に捉えられなかった．

そこで本システムでレポートを採点することで，作文スキルに関する部分の採点の正確性を保ち，短時間で作業を終えることができた．また，以降の学生指導に役立つ情報をまとめ，フィードバックを試みた．

6.3.3 教育実践による効果測定

6.3.3.1 実践内容

2016年度に実施したクラスCのレポート1および2を自動採点した．レポート1の結果をクラス全体にフィードバックし，授業最終回にレポート2の課題を提示し，変化を確認した．また前年度（2015年度）に実施した授業で提出されたクラスAおよびクラスBのレポート1を自動採点し，クラスCの同テーマのレポート1とともに，クラスの傾向などを分析した．対象となるクラスの文書数や平均文字数を表6.3.2に示す．

表 6.3.2: 採点したレポート

授業実施年度	クラス	文書数	平均文字数（標準偏差）	採点対象レポート
2015	A	43	427.6 (216.0)	レポート1
2015	B	40	325.5 (171.7)	レポート1
2016	C	42	443.0 (166.3)	レポート1
2016	C	42	438.1 (134.5)	レポート2

図 6.3.1 は採点の過程で表示される作文技術基礎統計情報の画面のキャプチャである．

ID	文体統一	誤字脱字	構文	主述関係	冗長さ	表記ゆれ	文字数	文章数	句読点数	漢字使用率	語彙水準
01	0	3	4	11	1	1	487	7	17	0.341	2.873
02	1	1	2	3	0	1	397	8	20	0.310	2.756
03	0	2	0	0	0	0	351	7	21	0.359	2.791

図 6.3.1: 採点時に表示される作文技術基礎統計情報

レポート 1 の自由記述文のテーマは次のとおりである．レポート 2 は課題提示としては同じであるが，データを自由に探す問題に変更してある．

> 「本シートの表は日本政府のデータカタログサイトで公開されている国土交通省 観光庁のオープンデータである．この表から『外国人旅行客に関する調査報告書』を作成することになった．どの部分をどのように利用してもよいので，ここからわかることを 1 つ自分なりに発見し，表とグラフを作成せよ．作成した表やグラフから言えることなど自分の見解を，少なくとも 200 文字以上記述すること.」

6.3.3.2 効果の確認

(1) 採点負担の軽減

本システムの利用により，文章作法や作文スキル等の添削および評価の時間的負担や評価のバラツキの軽減が可能となった．例えば 40 名の自動採点では，Style・Skill の採点及び総合成績レベル予測結果を得るまで数分足らずで行うことができる．アクティブ・ラーニングでは，授業実践に加え，厳正な評価と授業改善の取り組みが重要である．本採点システムを利用して時間を短縮することで，レポートの質向上に向けての論点やテーマを明らかにして議論するなど，教員と学生は双方向に意義ある内容に時間を費やすことができる．

(2) クラス毎の傾向の把握

同じテーマのレポート 1 をクラス別に処理し 3 クラスで比較したところ，作文スキルや漢字の利用などに違いがあることがわかった．例えば 6.3.2 は漢字の使用率をクラス別にグラフ化したものである．クラス B は A，C よりも漢字資料率が高い学生が多い．

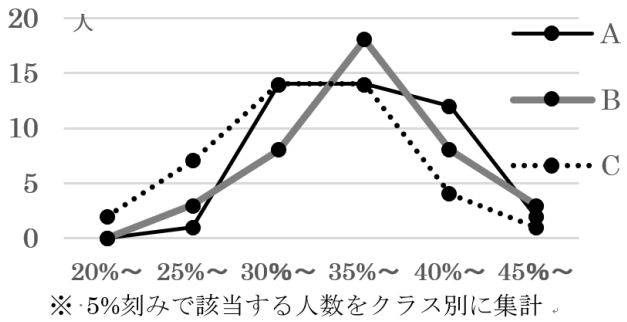

図 6.3.2: クラス別漢字使用率

また図 6.3.3 のレポートのエラー数の状況は，学生毎の文章校正エラーの数をプロットしたものである．

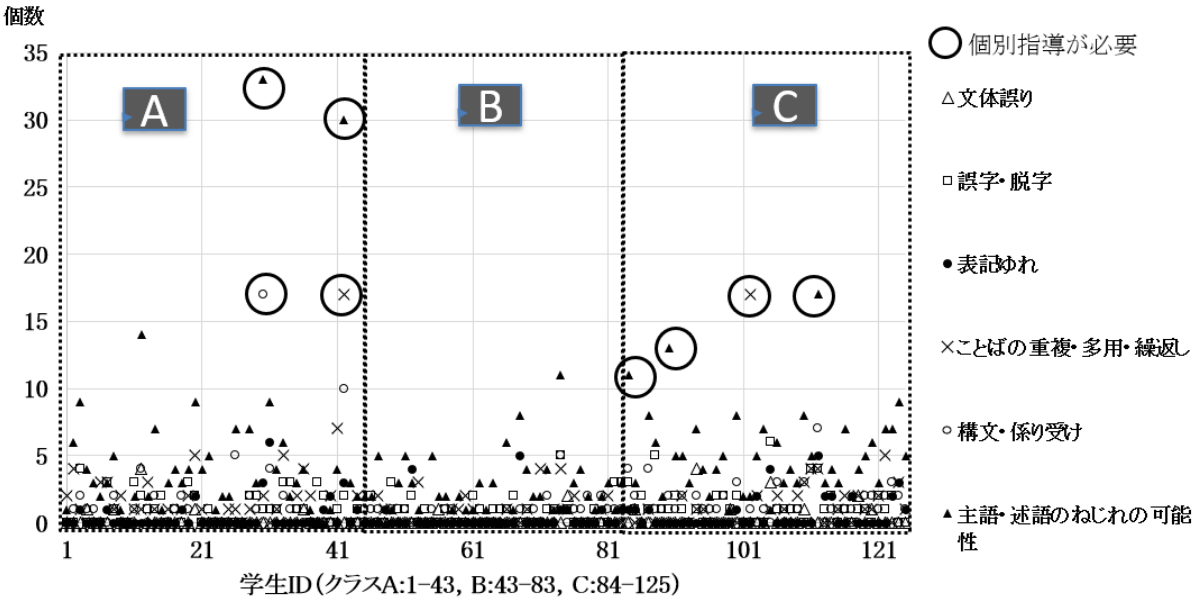

図 6.3.3: レポート 1 のエラー数の状況

ここでもやはり，クラス B は全般にエラーが少ない．授業内での説明や，レポートの課題提示の際，A，C クラスに対しては，文章校正に関する話を詳しく伝える必要がある．

(3) 作文スキルの向上

図 6.3.3 より，クラス A やクラス C は，▲や○の外れ値があり，個別指導が必要であることがわかる．そこで図 6.3.4 のように，該当学生にのみ Moodle でコメントを返し，改善指導を行った．

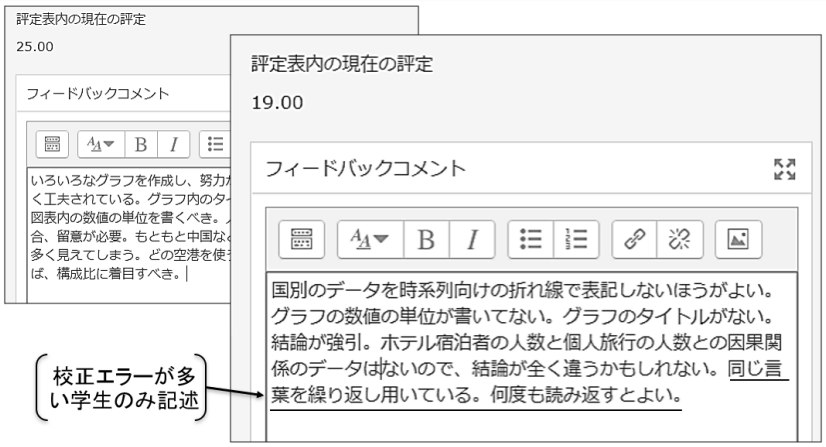

図 6.3.4: 個別指導コメントの例

次にクラス C について，個別指導が必要であった学生（図 6.3.3 の○で示した 4 名）のレポート 1 とレポート 2 を採点し，統計情報を確認すると，表 6.3.3 のとおりエラーの個数が「誤字脱字」と「冗長さ」（下線付き）を除き，すべて減少していた．

表 6.3.3: 学生の個別指導後のエラー数の変化

ID	文体統一	誤字脱字	構文	主述関係	冗長さ	表記ゆれ	レポート 1 文字数	レポート 2 文字数	文字数増減
84	1	-1	-2	-6	3	-1	487	576	89
90	0	-1	-1	-9	0	0	324	347	23
102	0	0	0	4	-15	0	495	471	-24
112	0	-4	-4	-13	-1	-2	1112	433	-679

また，表 6.3.4 のクラス C の語彙力の変化のとおり，漢字使用率や，語彙力，語彙の豊富さ（トークン比）が僅かに上がり，標準偏差が下がっており，改善が確認できた．これらは本授業での取り組みが直接の要因とは特定できないが，レポート 2 の提出まで 1 か月足らずであることから，ある程度の影響があったと推測できる．

表 6.3.4: クラス C の語彙力の変化

レポート（授業回）	平均文字数（標準偏差）	漢字使用率 (適正値 35%以上)	語彙水準 (最高点 6.0)	語彙の豊富さ (最高点 1.0)
レポート 1 (10 回)	443.0 (166.3)	34.0%	2.79	0.61
レポート 2 (15 回)	438.1 (134.5)	37.0%* (+3%)	2.95* (+0.16)	0.62 (+0.01)

※ウィルコクソンの順位和検定による．　n=42,　*： p ＜ 0.05

(4) 学生のモチベーション

随時行う授業評価では，授業の満足度のクラス平均は 4.5 以上，学生のやる気度は 5.0 前後の数値を獲得している（何れも 0-6 段階評価）．

6.4　レポート作成時の校正　ー学生の利用事例ー

6.4.1　実践した授業概要と科目の位置づけ

学生へのフィードバックに関する検証を行うために，2017 年度に担当した授業で，システムを試用した結果を報告する．取り組みを行った授業は，『ICT とグローバルコミュニティ』（2〜4 年次対象 1 期開講 2 単位，受講者 26 名）である．国際教養学科の専門科目ではあるが，学部・学科を問わず受講でき，一般の教養科目と同等に位置づけられる．

6.4.2　実践内容と効果

第 12 回授業で，次のようなレポートの課題を提示した．

「ＩＣＴの発展がもたらす社会問題を一つ取り上げ，原因や対策など 2000 字以上で論じなさい」

第 14 回授業までにテーマを決め，400 字以上で「概要」をまとめるよう指示し，この「概要」を文書校正の練習に用いた．学生は授業内で，システムにレポートの「概要」をアップ

ロードして提出し，チェックされた内容のエラーメッセージを確認後，訂正して再度アップロードする作業を 15 分程度行った．

以下は，学生 A のレポートから一文を取り出し，その一部を掲載したものである．文長 203，読点 5 つである．

「ICT による個人的な問題を自分の周りで何か起こっていないかと考えていたところ，最近頻繁に迷惑メールが届くことがあり，そのメールの発信源はどこなのか，‥‥（省略）‥‥ について述べることにした。」

チェックのためにアップロードし，返されたメッセージの一部を，図 6.4.1 に示す.「言葉の重複」,「一文が長すぎる」,「同じ助詞の繰り返し」など，訂正すべき個所が指摘される．なおこれらは，日本語表示も可能である．

```
DoubledWord: Found repeated word "ところ".
DoubledWord: Found repeated word "という".
DoubledWord: Found repeated word "迷惑メール".
SentenceLength: The length of the sentence (203) exceeds the maximum of 120.
CommaNumber: The number of commas (5) exceeds the maximum of 3.
DoubledJoshi: Found repeated Joshi word "が"
DoubledJoshi: Found repeated Joshi word "の"
DoubledJoshi: Found repeated Joshi word "を"
```

図 6.4.1: 文章校正メッセージの例

終了後に，利用した感想として所感を自由記述文で書いてもらった．表 6.4.1 に，所感データの基礎情報を示す．

表 6.4.1: 学生所感の基礎情報

回答者数	最大	最小	平均	標準偏差
25	48	202	102.5	48.1

受講者 26 名中 25 名 (約 96%）から得たコメントを，システムに対する肯定／否定から分けると，図 6.4.2 となる．概ね肯定的意見であった．また，気づきを書いている学生は 19 名（76.0%）であった．図 6.4.3 は，所感の頻出語 (10 回以上出現した語のみ抽出）のネットワーク図[2]である．丸が大きいほど，出現頻度が高いことを示す.「便利だと思う」や「自分の文章が長い」などの気づきがわかる．

[2]簡易型テキストマイニングプラグイン TeMP[1] で処理

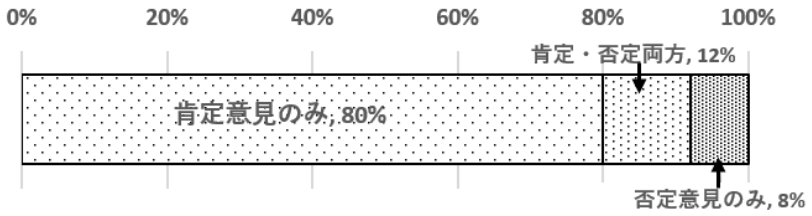

図 6.4.2: 利用後の所感の内容

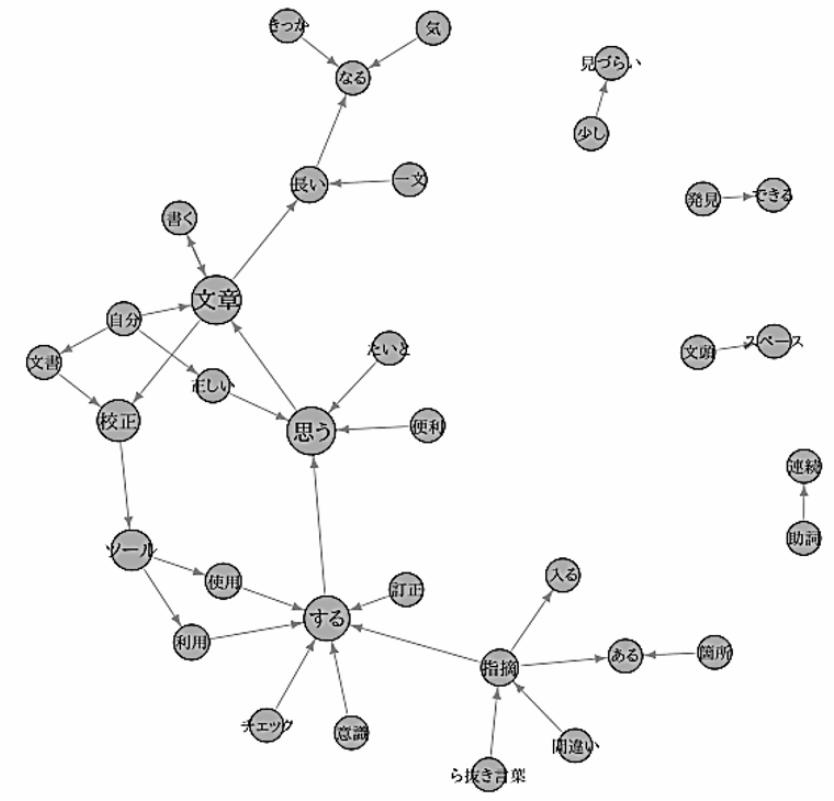

図 6.4.3: 所感の頻出語のネットワーク図

これらの具体的なコメントの例を，表 6.4.2 に抜粋する．

表 6.4.2: 学生所感の内容

肯定的意見や気づき	一文の中に記入する「に」や「の」の多さが目立った。個人的には作文能力に自信があったため，こういった盲点を発見できたことは有効であった。
	自分ではなかなか気づかくことができない助詞の重複や一文の文字数を文章校正で直すことができてよかった。これからもこのような機会があれば利用しようと思った。
	自分の文章，単語の使い方の傾向が分かったような気がしました。読みやすい文章を書くために，メリハリをつけてレポートを書きたいと思いました。
肯定と否定	自分では見つけられない間違いを発見できる点でよいツールだと思った。でも，校正ツールの正確性が問われた。
	表現が難しくて理解するのに少し時間がかかった。また，自分の文章がいかに綺麗じゃないかがよく分かった。
否定的意見	まず初めに思ったことは，構成の訂正の部分が見にくいと思いました。様々な間違いが指摘されているのに，そこが分かりにくいと思ったことが正直なところです。

6.5 むすび

AES 支援システムを導入しレポートを採点することで，教員は，論作文スキルに関する評価の厳正化を保ち，時間的な負担を軽減できる．また統計情報を分析することで，以降の授業運営に役立つ情報を得ることができる．レポートの多くは授業終盤での提出となる．今後は授業所感の自由記述文など，随時記述する内容を採点し分析することで，早期にクラスの傾向や課題を見つけ出すことへの利用が考えられる

一方学生は，事前にレポートをチェックするなど，最終的な提出を行うまでの主体的な取り組みが可能である．今後はフィードバックのユーザインターフェースを整え，実施する予定である．

第7章　自動採点精度向上に向けた語彙レベル辞書の構築

本章では，自動採点の精度向上に向けた日本語語彙レベル辞書の構築について述べる[1]．以下 7.1 節で本章の概要を，7.2 節では提案済みの AES 支援システムにおける語彙水準の計算方法と問題点について述べる．7.3 節で，提案する難易度計算に関する理論的な枠組みを説明する．7.4 節で，構築手順や語彙レベル（難易度）の具体的な計算方法を述べる．7.5 節で，提案モデルに従って語彙レベル辞書を作成し，7.6 節で，作成した語彙レベル辞書を用いた実験結果を報告する．7.7 節でまとめと今後について言及する．

7.1　はじめに

構築した AES 支援システムの採点評価項目の 1 つである語彙水準について，砂川らの日本語教育語彙表の難易度を用いているが，大学生のレポートで使用される単語が十分に網羅されていないため，採点精度に課題がある．そこで語彙水準算出に関係する評価値の精度向上を目的とした，網羅性の高い日本語語彙レベル辞書の構築手法を提案する．まず，大学生のレポートに出現する広範な単語を網羅するコーパスに，トピックモデル（LDA）を適用し，出現確率を指標とした難易度を算出する．希少性が高く出現確率を求めることが困難な単語は，従来から利用されている単語重要度 TF-IDF 値を出現確率の代替値とすることで難易度を補完し，語彙レベル辞書を構築する．基礎教育授業の学生レポートデータを用いた評価実験により，単語の採点漏れが解消できることを確認した．また，語彙水準の採点項目に関して，手動採点による評価値との相関を確認した結果，4.9%の精度向上が認められた．

7.2　AES 支援システムにおける語彙水準評価項目の計算方法と問題点

第 3 章 3.3.1 に示すとおり，5 つの評価観点を持つレポート採点用ルーブリックと細分化した 25 評価項目を提案し，Style と Skill を中心に自動採点システムを構築している．「V.Skill」の細分化した 4 つの評価項目の採点基準は，表 7.2.1 のとおりである．プロトタイプシステムで採点精度を確認したところ，評価観点「V.Skill」の手動採点結果と自動採点結果との相関

[1]本章は, 文献 [77] [79] [80] を加筆・訂正したものである．

が 0.255 であった (表 5.3.1)．採点処理を見直したところ，評価項目「25) 語彙の水準」に改善すべき点が見つかった．

表 7.2.1: Skill の自動採点用評価項目の採点基準

評価項目	採点基準
22）漢字の使用率	文書全体で 32%以上が望ましいとする
23）文長の妥当性	平均文長 26〜41 文字を適正範囲とする
24）語彙の豊富さ	異なり語数／述べ語数 により判断する
25）語彙の水準	レポート内の主要語彙の難易度平均値で判断する

「25) 語彙の水準」では，レポート（文書）内で使用される主要語彙（内容語と言われる名詞，動詞，形容詞，副詞）の平均語彙水準を求めている．平均語彙水準を L，主要語彙となる単語の異なり語数を M，単語 t の難易度を $DL(t)$，m 番目の単語 t の重みを $w(t_m)$ とすると，ある文書 d の平均語彙水準は式（7.2.1）で表すことができる．

$$L(d) = \frac{1}{M} \sum_{m=1}^{M} (DL(t) \cdot w(t_m)) \tag{7.2.1}$$

ここで難易度 $DL(t)$ は，日本語教育語彙表で単語毎に設定されている語彙の難易度 1〜6 を用いている．また各単語の重み $w(t)$ は，各文書内における出現頻度に基づき計算する．

例えば，次のような記述文からなる文書の語彙水準を求める場合，下線部の主要語彙の難易度平均を，語彙水準として算出する．

「大国 アメリカは日本からの資金調達が必要ないのではないかと考察する.」

これら 6 つの単語の日本語教育語彙表における語彙の難易度は表 7.2.2 のとおりである.「資金調達」は日本語教育語彙表に存在しないため計算されない．したがってこの文書の平均語彙水準は，式（7.2.1）より，次のように計算される．

$$(5 \times\ 7.4 + 1 \times\ 4.1 + 1 \times\ 6.4 + 3 \times\ 4.2 + 5 \times\ 17.5)/5 = 29.52$$

なお，表 7.2.2 の各単語の重みは，上記文書が実験用学生レポートデータ（インバウンドをテーマにした小レポートの文書集合）内の文書であり，全文書中の単語の出現頻度をもとに算出した単語重要度である．

表 7.2.2: 語彙水準の計算要素の例

語彙の難易度	5	1	1	—	3	5
単語	大国	アメリカ	日本	資金調達	必要	考察
重み	7.4	4.1	6.4	7.4	4.2	17.5

表 7.2.3 は，日本語教育語彙表の 1〜6 の難易度毎に存在する単語数をまとめたものである．「6. 上級後半」の単語数は少ない．

表 7.2.3: 日本語教育語彙表の難易度別単語数

語彙の難易度	単語数
1. 初級前半	424
2. 初級後半	792
3. 中級前半	2,300
4. 中級後半	6,465
5. 上級前半	6,379
6. 上級後半	1,560
合計	17,920

実際，実験用学生レポートデータの採点では,「アベノミクス」「食文化」,「家電量販店」,「無形文化遺産」,「民泊」など，中級以上と推測できる語が含まれておらず，語彙水準の採点対象から漏れていた．

第 2 章の表 2.3.1 で示した他の既存の語彙表のうち,『現代日本語書き言葉均衡コーパス』短単位語彙表は大学生のレポートに出現する単語を十分網羅していない．また均衡コーパスであるため，実際の利用頻度を十分に反映していない．特にレポート採点では，難解な単語を使う方が得点が高くなるとは必ずしも言えないため，現状に則した利用頻度を用いる方が望ましい．Simple PPDB : Japanese は，非均衡コーパス Wikipedia を利用しているものの，語彙の難易度を 3 段階に集約しており，学生間で差が生じにくく語彙水準の採点精度が期待で

きない．そこで，大学生のレポートで使用される単語を網羅する語彙レベル辞書の構築手法を提案し，日本語教育語彙表に替えて採点精度の向上を目指す．

ここで語彙レベル辞書に設定する難易度として，次の 2 つが考えられる．

- 日本語教育語彙表に倣い，追加する単語に難易度 1〜6 に則した離散値を設定する
- 追加の単語だけでなく，日本語教育語彙表の既存単語も含め，別の値を難易度として振り直す

前述したように，日本語教育語彙表は難易度が高い単語が少ない．また，あらたに追加する単語によっては，難易度 6 ではなく，7 以上を設定すべき可能性もある．仮に 1〜6 に加えて難易度 7 を設定する場合，日本語教育語彙表の中に参考とすべき単語がないため，信頼性のある難易度設定が困難である．一方すべての値を新たに設定し直すには多くの時間を費やす．そこで，日本語教育語彙表に存在する単語の難易度について相関を保つような値を計算で求め，あらたに設定する．

7.3 語彙の難易度計算のための指標

本節では，構築する語彙レベル辞書に設定する語彙の難易度を求めるにあたり，単語の出現頻度ではなく出現確率を指標とする理由と，トピックモデルについて説明する．7.3.1 では語彙レベル辞書の目的と単語難易度計算の理論的枠組みを，7.3.2 ではトピックモデルについて述べる．

7.3.1 語彙レベル辞書構築の目的と難易度指標の理論的枠組み

語彙レベル辞書は，採点用プログラム群の中にテーブルとして保持され，採点プログラムから随時参照される．項目は，ID（単にシーケンス No を振ったもの），TERM（表記），POS1（品詞大分類），POS2（品詞中分類），DL（日本語教育語彙表の語彙の難易度．但し存在するもののみ），D（あらたに求める単語難易度）とする．D はレポートの語彙水準値を求める際の難易度として，これまで計算に用いていた日本語教育語彙表の語彙の難易度 DL に替えて利用するものである．TERM の数が多く網羅性が高いこと，採点に利用する D が現状に相応しい難易度（値）であることが，語彙レベル辞書構築の主たる目的である．

前者については，現時点で入手可能な大規模コーパスで大学生レベルが頻繁に利用する単語を網羅し，新語にも対応できる Wikipedelia を用いることが，1 つの解決策であると考え

る．内容が随時更新され，配布されるコーパスのデータベースが定期的に更新されるが，提案する構築手順にしたがえば，随時再構築可能である．

後者の難易度の計算については，難易度算出指標として単語の出現確率を利用することで可能である．先行研究では単語の難易度を，単語親密度，出現頻度，単語重要度から求めることが一般的である．これらの指標は何れも，出現頻度を重要視する考え方である．一定の範囲内で出現頻度が高ければ，よく見る単語であり親密度が増し，平易な単語であることを意味する．専門性が高くなり一般に目にしない，あるいは使用しない単語は，難易度が高いことになる．しかしながら，単語の出現頻度は，トピック（話題）によって異なる．例えば，「観光地」という単語の出現頻度は，観光に関するトピックと経済に関するトピックでは，双方に出現する可能性があるが，出現頻度が異なる．したがって，各単語がどの程度一般性があり，どの程度専門性が高いかを推し量って単語の難易度指標とするには，トピックごとの出現確率を算入すべきである．

7.3.2 トピックモデル

トピックモデルは，文書を生成するための確率モデルである [81]．1 つの文書が複数のトピックを持つと考える．各文書には潜在的にトピックが含まれており，そのトピックが含まれる確率や，さらに各トピック内に含まれる単語およびその確率も潜在的に決まるとする．生成された文書は，文書のテーマに関連するトピックの出現確率，および，各トピック内の単語の出現確率により選ばれた単語の集合になる．図 7.3.1 は，学生のレポート生成過程をイメージしたものである．例えば「日本の観光政策についてレポートを作成しなさい」という論題 (テーマ）が課せられた場合，与えられた論題から複数のトピックが浮かぶであろう．さらに各トピック内で浮かぶ単語が複数あり，それらの単語が共起して一つのレポートを生成するとする．トピックや単語を思い浮かべる確率が，出現確率を意味する．

例えば Wikipedia コーパスから単語の出現確率を求めて語彙レベル辞書を作成する際，単純な出現確率は次式で求まる．

$$P(w) = \frac{\text{単語 } w \text{ の出現総数}}{\text{ウィキペディア全体の単語総数}}$$

この場合は Wikipedia の全体を 1 つのトピックとして求めることになる．しかし実際は複数のトピックで構成される．そこで，単語の出現頻度を現状に沿った，より正確な値を求めるために，LDA を採用する．LDA は文書中の単語の順番は無視して，単語の共起関係に着

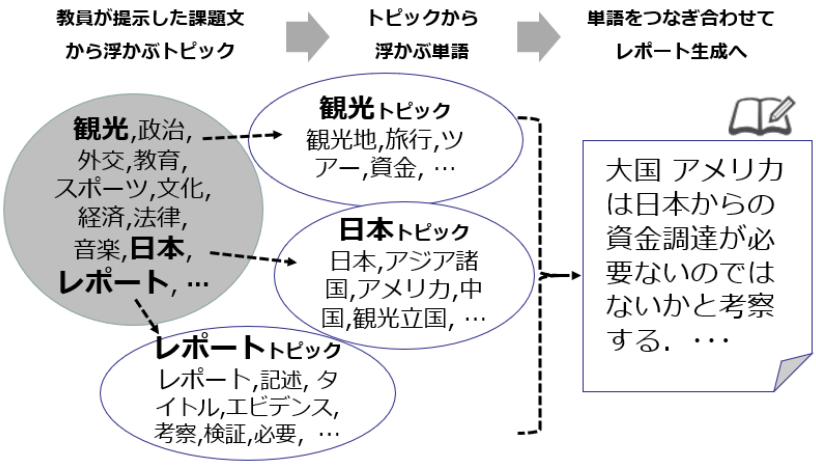

図 7.3.1: 学生レポート生成過程

目する確率モデルである [82]．文書に含まれる単語とその確率のみで文書全体をベクトルで表すことが可能であり，大規模なデータの処理に向いている．

図 7.3.2 は，文献 [81] で提示されている LDA のグラフィカルモデルである．図中の M は文書数，N は文書内の単語数を示す．○（α，β，θ，z）は未知の潜在変数である．●（w）は，文書中のある単語を示す．矢印は影響を及ぼす関係を示している．例えば Wikipedia 全体を文書集合とする場合，M は，Wikipedia の全記事数を示し，N はある記事内の全単語数，w はその記事内の 1 単語を示す．この単語が記事を生成する際含まれる（選ばれる）ためには，この単語を含むトピック z 内での潜在的な出現確率，およびそのトピックが Wikipedia 全体で出現する潜在的な確率 θ が影響して決定される．したがって，各単語 w の出現確率は，LDA による文書の生成過程で求まる．Python のライブラリ gensim は，LDA の文書生成モデルを実現する．トピック数をパラメータとして与えると，与えられた単語 w の集合（BoW : Bag of Words) から，各単語の出現確率やトピックの出現確率を出力する．これにより単語毎の出現確率を求めることが可能となる．

7.4 語彙レベル辞書構築方法の提案

本節では網羅性が高い語彙レベル辞書の構築方法を提案する．7.4.1 では，具体的な構築手順を，7.4.2 で，語彙レベル辞書に設定する単語難易度の計算方法を述べる．7.4.3 で，希少性が高く出現確率を求めることが困難な場合の補完方法を提案する．

[2]出典：文献 [81] の Figure 1

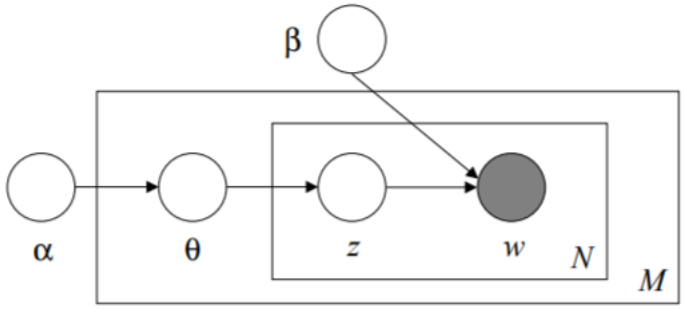

図 7.3.2: LDA のグラフィカルモデル[2]

7.4.1 語彙レベル辞書の構築手順

図 7.4.1 に語彙レベル辞書の作成手順を示す．はじめにもととなるコーパスを取得し，データを整形する（処理 A）．例えば，日本語 Wikipedia コーパスを用いる場合は，ダウンロードしたデータから有効な記事本文を抜粋する．コーパスによって，必要となる整形は異なるが，不要なデータを除去し，文章あるいは単語の集合とする．対象とする品詞は，内容語である名詞，動詞，形容詞，副詞とする．名詞は文章の内容を表現する意味語としての役割を持ち，語彙力を顕著に示すため，文章の重要度や難易度，類似度を測定する研究の多くで取り上げられている．本研究は，学生レポートの採点に用いる語彙レベル辞書構築を目的としているため，動詞，形容詞，副詞も含める．その他，半角・全角の統一，ストップワード除去，未知語除去を行う．

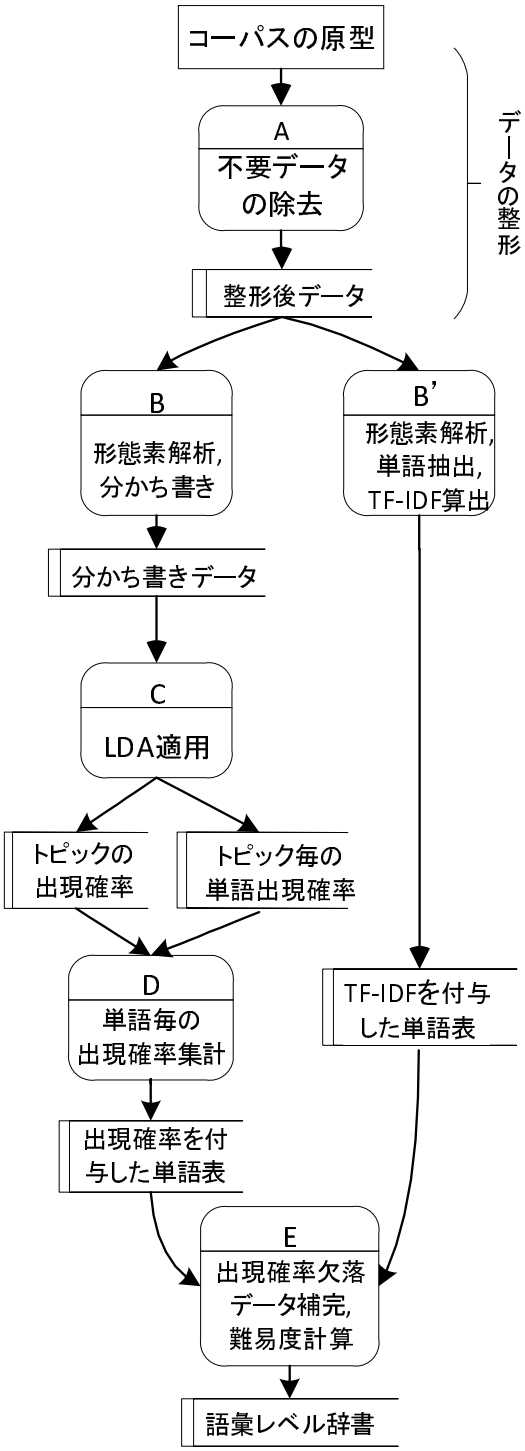

図 7.4.1: 語彙レベル辞書の作成手順

図 1 の左側の流れ（処理 B〜D）は，LDA の適用により出現確率を求める処理である．処理 B で MeCab により形態素解析を行い分かち書きしたデータに，処理 C で LDA を適用する．処理 D では，処理 C で得られたトピックの出現確率と，トピックごとの単語の出現確率を掛け合わせ，単語毎に出現確率を集計する．

右側の流れ（処理 B'）は，語彙レベル辞書の原型となる語彙表の作成と，出現確率の算出が困難なデータに対応するための処理である．B と同様に形態素解析を行い，語彙表を作成し，品詞などの必要な情報や TF-IDF を求めて付与する．

処理 E では，先の処理 D の結果と処理 B' の結果をマッチングし，単語出現確率データが存在する場合は単語出現確率を，存在しない場合は処理 B' で得られた TF-IDF 値を用いて代替値を求める．最後に単語出現確率または代替値を指標として単語の難易度を計算し，語彙レベル辞書を完成する．以降で，処理 C・D・E について詳述する．

7.4.2 LDA によるトピックモデルの適用と難易度計算方法

コーパスに LDA を適用すると，トピックごとに，単語のコーパス内での出現確率を求めることができる．2.3 節で述べたように，近年，単語の出現頻度を素性に，単語重要度や単語親密度などの難易度指標を求める研究がしばしば見られるが，本研究では 7.3.1 項で説明した通り，出現確率を素性とする．

処理 C で，分かち書きデータに LDA を適用する．この時，トピック数を設定する必要がある．単語がトピックごとに矛盾なく分類された上で，できるだけ多くの単語の出現確率を求めることができるような数を設定する必要がある．この値はコーパスにより異なるため，最適なトピック数を事前に探索しておく必要がある．また各単語の出現確率と，各トピックの出現確率をそれぞれ求め，両者を掛け合わせて単語毎の出現確率を集計する．今，ある単語 t のコーパス全体での出現確率を $P(t)$ とし，トピック数 K，LDA により求められたあるトピック T のコーパス全体での出現確率を λ_k，あるトピック T での単語 t の出現確率を $p(t|Tk)$ とする．各単語は複数のトピックに関わる可能性があるため，すべての出現確率を合計する．このとき，各単語の出現確率は，式（7.4.1）で求めることができる．

$$P(t) = \sum_{k=1}^{K} \lambda_k \big(p(t|T_k)\big) \tag{7.4.1}$$

これらの出現確率 $P(t)$ を単語難易度指標とする．

さらに，語彙レベル辞書に設定する単語難易度を $D(t)$ とすると，採点では値が低いほど (出

現確率が低いほど）得点を高くしたいこと，単語難易度の上限を抑える必要があることから，式（7.4.2）で求めることとする．

$$D(t) = -\log_2 P(t) + \epsilon \tag{7.4.2}$$

7.4.3 出現確率算出の補完法

希少性の高い単語については，式（4.1）の出現確率 $p(t|T_k)$ が限りなく 0 に近く，求めることが困難な場合がある．そこで，単語重要度 TF-IDF で代替する補完法を提案する．出現確率は，もともと単語の出現頻度を素性としているので，両者に相関関係が認められる．また TF-IDF も，出現頻度を素性としており，単語難易度指標を求める方法の一つとして先行研究でも用いられている [23]．今，ある文書 d に出現する単語 t の出現頻度を $tf(t,d)$，全文書数を N，単語 t が出現する文書数を $df(t)$ とする．ある単語 t のコーパス C での単語重要度 TF-IDF を $w(t,C)$ とすると，TF-IDF は式（4.3）で求められる．

$$w(t,C) = tf(t,d) \cdot \log \frac{N}{df(t)} \tag{7.4.3}$$

次に，予め出現確率を求めるための単回帰式を，出現確率と TF-IDF の相関関係から求めておく．この式の係数を α ，β とすると，出現確率 $P(t)$ の代替値 $P'(t)$ は式（4.4）で求めることができる．

$$P'(t) = \alpha \cdot w(t,C) + \beta \tag{7.4.4}$$

よって，LDA により出現確率が求まらない単語については，式（4.2）の $P(t)$ を，$P'(t)$ に置き換えて求める．

7.5 語彙レベル辞書の構築

本節では，前節の提案にしたがって，いくつかの実験を交えながら，語彙レベル辞書を構築する．7.5.1 では，もととなるコーパスの選定と整形を，7.5.2 では，トピック数の探索とト

ピックモデルの適用結果を述べる．7.5.3 では，出現確率を求めることが困難な単語の補完を，7.5.4 では，語彙レベル辞書を構築した結果を述べる．

7.5.1 コーパスの整形

Wikipedia には 2017 年 12 月時点で 3,225,450 件の記事が存在し，年々，追加・更新されているため，網羅性の高い語彙レベル辞書の作成が可能である．関連記事が互いにリンク付けされていること，読み手は興味があり，自分のレベルにあった記事を読むことから，Wikipedia の内容はレポート作成や採点時に影響する．また，大勢が執筆に関わり，読みやすく理解しやすい単語など，より的確な表現が多く使われ，使用すべき単語の適性を示す出現頻度となっている．本研究では単語の採点漏れを防ぐ目的が第一であるため，大規模な Wikipedia コーパスをもとにして，あらたに語彙レベル辞書を構築する．

日本語 Wikipedia データベースサイト[3]より，全記事データをまとめた XML ファイルをダウンロードし展開する．2017 年 12 月 25 日時点の最新版である記事のタイトル数として 3,225,450 件が登録されており，日本語本文を含む有効記事 756,666 件を処理する．形態素解析器 MeCab を利用して分かち書きを行う．新語に対応するために，Wikipedia の全記事のタイトルと，はてなキーワードからユーザー辞書を作成し事前に追加登録済みである．対象とする品詞（名詞，動詞，形容詞，副詞）を抽出する．その他，半角・全角の統一，ストップワード除去，未知語除去を行う．

7.5.2 トピック数の探索

Wikipedia のデータ量は膨大であり，全体のトピック構成数の推定は困難である．トピックと似た「カテゴリー」と呼ばれるメタデータが各記事に付与されているが，カテゴリー数 218,191 を，トピック数として設定することは適切でない．また論文 [83] では，トピック数の設定や，希少性が高く出現確率の計算が困難な単語に関する対応など，具体的な辞書作成方法については触れられていない．一般的には，100〜数百のトピック数が用いられており，岩田は階層ディリクレ過程によるトピック数の推定方法を紹介している [84]．松河らはトピック数の値を変化させながら perplexity を求め最も低くなるトピック数を選定している [85]．その他，高い値のトピック数を設定して処理した後，トピック間の類似度で結合し最終決定する方法もある [86]．本研究では，日本語教育語彙表の各単語の難易度を踏まえた上で，新たに難易度を設定した語彙レベル辞書を構築する．また，難易度の高い単語の採点漏れを防ぐことが目的であるため，多くの単語の出現確率の算出を要する．さらに，新たに求める難易

[3] https://dumps.wikimedia.org/jawiki/

度と，日本語教育語彙表の難易度との相関が認められることが望ましい．そこで，100 トピックから探索的に LDA を適用し出現確率を確認することとする．その結果，表 7.5.1 に示すように，300 トピック以降は出現単語数が多いが，500 をピークに減少すること，日本語教育語彙表の語彙の難易度との相関は，トピック数による大きな差異は認められないことが確認できた．

表 7.5.1: トピック数の探索

トピック数	200	300	400	500	600	700
相関 [1]	-0.210	-0.225	-0.237	-0.245	-0.211	-0.214
出現確率算出可能単語数	7,030	10,188	11,822	14,867	13,385	12,794

[1] 日本語教育語彙表および実験用学生レポートデータのうち，全てのトピック数で出現確率が存在する単語 1,857 件について，日本語教育語彙表の単語難易度レベルと出現確率とのピアソンの相関を求めた．

図 7.5.1 は，python library の gensim を用いて，トピック数 500 として Wikipedia データに LDA を適用した結果の一部の抜粋である．トピック ID:0 と ID:323 の出現確率上位 10 の単語とその確率をペアで示している．1 行目（ID：0）は，国家，政治，国民，社会などの単語が集まっており，政治に関するトピックと推測できる．2 行目（ID：323）は，観光に関するトピックと推測できる．出力された 500 トピックを確認し，何れも意味のあるまとまりと判断できたことより，500 トピックを採用する．


(0, [('国家nation', 0.075782895), ('政治politics', 0.043928403), ('国民people', 0.027891846), ('社会society', 0.024236703), ('改革reform', 0.022163419), ('政治的political', 0.019293314), ('政策policy', 0.018285373), ('時代era', 0.010872778), ('権力power', 0.009812207), ('国country', 0.0093505923), ('民主democracy', 0.0079780845), ・・・

:

(323, [('観光sightseeing', 0.12125151), ('旅行travel', 0.095982991), ('訪れるvisiting', 0.095198631), ('ツアーtours', 0.07273744), ('観光客tourists', 0.04697549), ('観光地sightseeing spots', 0.017037462), ('魅力attractions', 0.013554713), ('観光協会Tourism Association',

図 7.5.1: LDA 適用結果の例

表 7.5.2 は，提案した手法で日本語教育語彙表に存在する単語の Wikipedia 内の出現確率を求め，難易度毎に平均を算出した結果である．難易度が高いほど，出現確率が低いことがわかる．出現頻度が低い単語は目にする機会が少ないためなじみがなく難易度が高いと言える．したがって算出した結果が妥当であることが確認できる．なおここでは，実験用学生レポートデータ内のすべての単語の集合をキーとして，Wikipedia 全体から抽出した記事（全体の約 36%に相当）で調査したものである．

表 7.5.2: 単語の出現確率の難易度別平均値

日本語教育 語彙表難易度	単語数	単語毎の出現確率 総和の平均
1	302	0.0901
2	607	0.0825
3	1,783	0.0581
4	4,850	0.0396
5	4,448	0.0258
6	871	0.0189
平均	2,144	0.0392

7.5.3 出現確率データ補完値の計算

出現確率 $P(t)$ を求めることが困難な単語について，TF-IDF で補完するために，両者の相関関係から回帰式を求める．図 7.5.2 は実験用学生レポートデータに出現する単語のうち出現確率を求めることができる 650 語について，$P(t)$ と TF-IDF の相関関係を散布図で表したものである．高い相関（0.834）が認められ，式 (7.5.1) の回帰式を得ることができる．したがって，代替値 $P'(t)$ は次式により求め，データを補完する．なお回帰式や相関係数の算出は，何れも R により求めたものである．

$$P'(t) = w(t, C) \cdot 2 \cdot 10^{-6} + 0.0215 \tag{7.5.1}$$

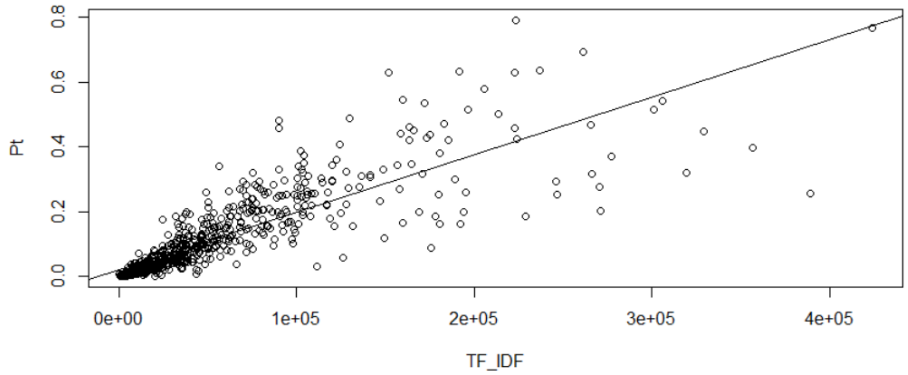

図 7.5.2: TF-IDF と $P(t)$ との相関

7.5.4　語彙レベル辞書

図 7.5.3 は，語彙レベル辞書の一部を抜粋したものである．7.2 節で採点漏れとして例示した「資金調達」の単語難易度が 4.48 と設定されていることがわかる．

ID	TERM	POS1	POS2	DL	D
419	思い	名詞	一般	3	3.10
420	施設	名詞	一般	4	2.33
421	私	名詞	代名詞	4	2.43
422	私たち	名詞	一般		4.81
423	資金	名詞	一般	4	3.46
424	資金調達	名詞	一般		4.48
425	資料	名詞	一般	4	3.13
426	事前	名詞	一般	4	3.91
427	持つ	名詞	一般	2	1.62
428	時間	名詞	一般	1	2.64

図 7.5.3: 語彙レベル辞書の例

7.6　語彙レベル辞書の評価実験

本節では，構築した語彙レベル辞書を用いて，実験用学生レポートデータの語彙水準を計算した結果を述べる．第 5 章の表 5.2.1 と同じデータを用いて，成果を確認した．7.6.1 では，単語の網羅性に関する実験結果を報告し，7.6.2 は，採点精度と課題を述べる．

7.6.1　採点漏れの減少

83 件のレポートを形態素解析した結果，1,212 の異なり語を得た．記号等を除外すると，採点すべき名詞は 888 であった．このうち 230 の単語が，日本語教育語彙表に存在せず，採点対象外となっている．一方，構築した語彙レベル辞書では，表 7.6.1 に示すとおり，LDA 処理の時点で 166，TF-IDF 補完により 64 語を被覆することができ，最終的に全単語を採点対象とできた．

表 7.6.1: 採点漏れ率

参照する難易度	採点単語数	採点可能 単語数累計率	採点漏れ単語率
日本語教育 語彙表	658	658(74.1%)	230 (25.9%)
LDA による 追加採点	166	824(92.8%)	64 (7.2%)
TF-IDF に よる補完採点	64	888(100.0%)	0 (0.0%)

また学生レポートごとの採点漏れの変化を確認したところ，表 7.5.2 に示す通り，平均約 10 単語が新たに採点されるようになった．

表 7.6.2: 辞書変更による採点結果の変化

当初の 平均採点単語数	提案後の 平均採点単語数	平均増加数
26.76	36.65	9.89

採点対象となった具体的な単語の一部を表 7.6.3 に示す．

表 7.6.3: 採点対象となった単語の例

採点状況	個数	単語の例
提案難易度 (出現確率) により採点対象となった単語	166	インセンティブ，意外，円安，格安，気風，食文化，単身赴任，伝統文化，富裕層，民泊，無形文化遺産，利便性，歴然ほか
補完提案難易度 (TF-IDF から算出) により採点対象となった単語	64	アベノミクス，オープンデータ，家電量販店，爆買い，密航，免税店ほか

7.6.2 採点精度と考察

語彙水準の手動採点結果と比較したところ，構築した語彙レベル辞書による採点精度は，22.9%から 27.8%へと 4.9%向上した．さらなる精度向上に向けて，採点結果とレポートの内容を確認したところ，次のような知見が得られた．

(1) 出現確率が低く難易度が高い値となっている単語の中に，実際は難易度が低く使用頻度が低い単語が存在する（「あまり」「いろいろ」「たくさん」「ひとり」など）．

(2) 難易度が高い単語を多く使用していても，同様に難易度が低い単語が多いと，文書全体の語彙水準は低くなる．

（1）への対応として，すべて数字またはすべてひらがなで成り立っている単語の出現確率総和を，日本語教育語彙表の難易度 1 の平均値と同じ 0.0901（表 7.5.2）に置き換える．また，日本語教育語彙表で難易度が低いにもかかわらず提案難易度が高い単語について，より適正な難易度となるよう調整する．

（2）について精査するため，採点漏れが大きく改善された学生の記述文を確認した．表 7.6.4 は採点可能となった単語が多い文書を 2 つ取り上げ，それらの単語を抜粋したものである．

文書 a は語彙水準の順位が上がった例，文書 b は下がった例で，採点漏れが解消された単語を示している．文書 b のように，採点漏れが無くなったにも関わらず得点が下がる理由は，（1）で述べたように，平易な単語の漏れが少なからずあり，難易度が高い単語と同様に低い単語の多くが採点対象へとかわったためである．実際，日本語教育語彙表には，国名，地名

表 7.6.4: 採点結果の変化が大きい文書の事例

文書	採点単語数などの変動	採点対象となった単語 [1]
文書 a	採点単語数 +21， 語彙水準の順位 +3 (全文字数 512)	中国 (5)，あまり (2)，海外進出， 資金調達，地理的，漢字 文化圏，在日韓国人，朝鮮 戦争，戦火，朝鮮半島，密航， ダントツ，人面，これ (2)， その他
文書 b	採点単語数 +24， 語彙水準の順位 -15 (全文字数 1001)	訪日外国人，中国 (9)，台湾 (3)，香港 (4)，シンガポール， マレーシア，ドイツ，フランス， ロシア，その他，欧米 諸国

[1] （ ） 内の数値は，複数回出現した単語の個数を示す

などの固有名詞が一部しか設定されていないため，文書 b では多くの国名が追加採点されている．これらのほとんどは，難易度が低い．本ケースでは同じ単語を繰り返し使っている点で採点結果が下がるのは問題はないが，今後，語彙水準の計算式について，レポート内の難易度の分布を考慮した特徴量を導入する必要がある．

7.7 むすびと今後の課題

本稿では，レポート自動採点支援システムの採点項目のひとつである語彙水準の精度向上を目的として，Wikipedia コーパスから網羅性の高い日本語語彙レベル辞書の構築手法を提案した．難易度は，トピック内での利用頻度を加味し，Wikipedia 全体に LDA を適用して得られた出現確率を用いる．また，出現確率が希少な単語は TF-IDF で補完することで，網羅性が高い語彙レベル辞書を構築した．学生レポートデータを対象とした実験から，採点漏れを 100.0%近く解消することがわかった．また手動採点との相関から，語彙水準に関する採点精度は，4.9%の向上が認められた．今後は，次の 2 点を課題とする．

・単語の難易度について矛盾する値がないよう，構築した辞書を精査し，手動で調整する

・提案辞書を利用して採点する評価項目「25) 語彙水準」の計算式に難易度の分布を導入する

考察で述べたように，高レベルの単語が多い場合でも低レベルの単語が多いと，レポート全

体の語彙レベルのスコアが低くなる．そこで，高レベルの単語の文書内での分布を調べ， これを語彙水準の計算式に導入すればよい．高レベルの単語は日本語教育語彙表では難易度レベル 5 以上である．表 7.4.2 より，これに相当する単語は出現確率 0.0258 以下であるとして該当する単語の個数を調べ，文書全体の単語に占める割合を計算する．今，ある文書 d における，主要語彙となる単語の異なり語数を $M(d)$，出現確率 0.0258 以下の単語の個数を $Htn(d)$ とすると，式 (7.2.1) より，改善後の平均語彙水準 $L'(d)$ は，次式 (7.7.1) で得られる．

$$L'(d) = \frac{1}{M} \sum_{m=1}^{M} (DL(t) \cdot w(t_m)) \cdot \frac{Htn(d)}{M(d)} \tag{7.7.1}$$

今後は上記の考え方をシステムに取り込んで実験し，精度を確認する予定である．

第8章　終章

8.1　本研究の結論

本研究では，授業形式の教育現場において，採点負担の軽減，評価の厳正化，論作文指導支援を行う，教員・学生双方に役立つシステムの開発を目標に，ルーブリックに基づく自動採点支援システムを，Moodle プラグイン TeMP[87] を拡張し構築した．大学の初年次教育や教養教育など，基礎教育の授業で課す記述文や，エッセイタイプの 100〜2000 文字程度の日本語論作文を対象としている．文法や読みやすさ，語彙力，およびレポートの論題と記述文の類似度を自動採点し，論作文スキルにかかわるルーブリックの評価観点を予測する．また機械学習により，12 の評価項目から作成した分類器により総合評価を求め，教員に採点支援評価値として提示する．授業担当者がレポート採点を行う際に，自動採点結果をセカンドオピニオンとして参照する，あるいは文章作成スキル部分の評価に自動採点結果を利用することで，評価の厳正化や時間的負担軽減を図る．さらに，採点精度向上をめざして語彙水準評価で用いる語彙レベル辞書構築モデルを提案した．モデルにしたがって作成した辞書を用いて採点したところ，4.9%の精度向上が確認できた．

本システムを教育現場で試行したところ，教員・学生双方に一定の効果が確認できた．今後，分類のための学習データを増やして分類器を再作成し，より精度の高いものへと改善していくこと，学生へのフィードバックの内容を改善することで，有効利用が期待できる．また語彙レベル辞書の構築モデルは，コーパスに依存しないため，コーパスを専門性の高いものに変えることで，専門領域の採点に利用できる可能性が高い．

8.2　本研究の課題と展望

本研究の課題は，総合評価の採点精度を高めることである．そのためには，レポートの内容や論理性の採点，すなわち図 8.2.1 の「I .Content」や「II. Structure」の評価項目について，精度の高い自動採点を実現する必要がある．大学の基礎教育におけるレポートでは，論作文スキルと内容の是非との相関が確認でき，AES 支援システムのある程度の精度を見込める．しかしながら，I から V の配分は教員により異なるため，柔軟な対応ができる仕組みが必要であろう．また I ・II は，論題によるところが大きいので．今後さらに汎用化するため

には，内容や論理性の採点アルゴリズムの議論が必要である．

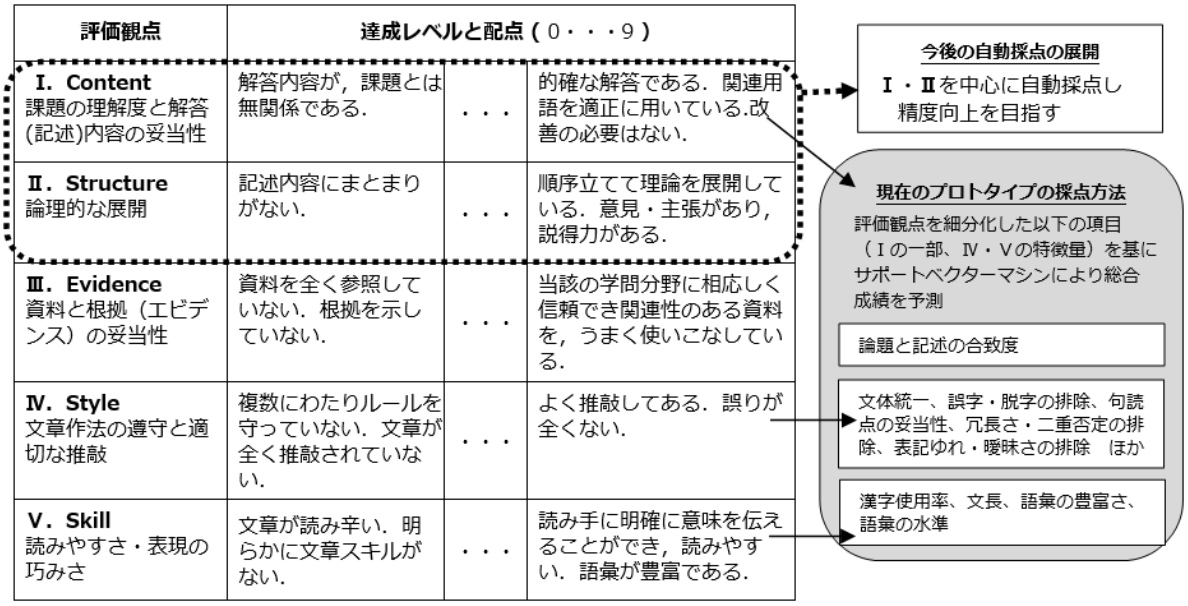

図 8.2.1: 精度向上に向けた今後の自動採点モデルの展開

合否を判定する試験の自動採点では，正解データを学習させるなど，事前に正答例やスコアが高い採点済みデータを投入し，各種の教師あり学習により，比較的高い精度で採点可能である．しかし，教育現場で採点するレポートは，科目や教員ごとに設問や目標達成度が異なり，多数の正答例を利用することが困難である．そこで教員が影響を受ける可能性が高い複数コーパスを利用することで，採点に用いる複数のニューラルネットワークモデルの構築を検討する．I については，レポートの論題やキーワード，シラバスなどのテキストデータからカテゴリを推測し，これらと同じカテゴリの単語ベクトルとの類似度を基に採点する．II については，理論的展開が高低どちらのレベルに分類されるかという，確率を基に採点を試みる．なお先行研究として，Kim ら (2014) や Ma ら (2015) は，畳み込みニューラルネットワーク（CNN）を用いた文章の分類を試みている [88] [89]．寺田ら (2016) は，CNN による句レベルの 2 値の分類（採点）が良好であることを [90]，Dong ら (2016) は，CNN により特徴量を自動的に導き出す方法を提案している [91]．こうした研究動向を参考に，内容や論理性部分の採点の手法を検討する．

謝辞

本研究の遂行にあたり，終始手厚いご指導・ご支援をいただいた，南山大学 理工学部 河野浩之教授に，心からお礼申し上げます．常に的確にご教示くださり，研究活動を続けることができました．研究内容はもとより，研究者としての心構えなど本質的な部分から，学際的な幅広い内容まで，実に多くのことを学ばせていただきました．あらためて感謝申し上げます．

本研究を審査くださいました，南山大学 奥村康行教授，沢田篤史教授，石原靖哲教授，名古屋大学 石川佳治教授には，多くのご助言を頂きました．心から感謝の意を表します．また，南山大学 大石泰章教授，鈴木敦夫教授にも大変お世話になりました．研究につながる多くのヒントを頂き，学問のつながりや楽しさを実感することができました．

本研究の基盤となるシステム TeMP の開発者である名古屋学芸大学梅村信夫教授には，システムの拡張を了解いただき，多くのご支援をいただきました．深く感謝申し上げます．

そして何よりも，本論文をまとめることができたのは，研究活動を理解し，常に支え励ましてくれた家族のおかげです．心から感謝いたします．本当にありがとうございました．

参考文献

[1] 梅村信夫, 山本恵. 教師用テキストマイニング・プラグインの開発と評価. *Proceedings of Moodle Moot Japan 2015 Annual Conference*, pp. 59–64, 2015.

[2] Terrel Rhodes. Assessing Outcomes and Improving Achievement: Tips and Tools for Using Rubrics. *Association of American Colleges and Universities*, 2010.

[3] ローネン・フェルドマン, ジェイムズ・サンガー. テキストマイニングハンドブック. 東京電機大学出版局, 2010. 辻井潤一 監訳, IBM 東京基礎研究所 訳.

[4] Semire Dikli. An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning and Assessment*, Vol. 4, No. 1, pp. 1–36, 2005.

[5] Mark D. Shermis, Jill Burstein, and Sharon Apel Bursky. Introduction to Automated Eessay Scoring. In *Handbook of Automated Essay Evaluation: Current Application and New Directions*, pp. 1–15. Routledge, 2013.

[6] Tsunenori Ishioka. Latest Trends in Automated Essay-Scoring Systems. *Kodo Keiryo-gaku (The Japanese Journal of Behaviormetrics)*, Vol. 31, No. 2, pp. 67–87, 2004.

[7] 黒橋禎夫. 自然言語処理. NHK 出版, 2015.

[8] Page and Ellis B. The Imminence of Grading Essays by Computer. *The Phi Delta Kappan*, Vol. 47, pp. 238–243, 1966.

[9] 西尾章治郎, 上林弥彦（編）. データベース. オーム社, 2000.

[10] 前川喜久雄（編）. コーパス入門. 朝倉書店, 2015.

[11] Yigal Attali and Jill Burstein. Automated Essay Scoring With e-rater? V.2. *ETS Research Report Series*, pp. i–21, 2005.

[12] 石岡恒憲. 作文テストにおけるコンピュータ利用と自動採点:—最新技術と今後の方向—. コンピュータ＆エデュケーション, Vol. 32, pp. 22–28, 2012.

[13] 石岡恒憲, 亀田雅之. コンピュータによる小論文の自動採点システム jess の試作. 計算機統計学, Vol. 16, No. 1, pp. 3–19, 2003.

[14] 石岡恒憲, 亀田雅之, 劉東岳. 人工知能を利用した短答式記述採点支援システムの開発 (言語理解とコミュニケーション) – (第 3 回自然言語処理シンポジウム). 電子情報通信学会技術研究報告, Vol. 116, No. 379, pp. 87–92, 2016.

[15] 石岡恒憲. 小論文およびエッセイの自動評価採点における研究動向 (<特集>テキストの自動評価). 人工知能学会誌, Vol. 23, No. 1, pp. 17–24, 2008.

[16] 石岡恒憲. コンピュータ上で実施する記述式試験について. Technical Report 19, (独) 大学入試センター／東京工業大学, 2016.

[17] Peter W. Fottz, Lynn A. Streeter, Karen E. Lochbaum, and Thomas K Landauer. Implementation and Applications of the Intelligent Essay Assessor. In *Handbook of Automated Essay Evaluation: Current Application and New Directions*, pp. 68–88. Routledge, 2013.

[18] Jill Burstein, Joel Tetreault, and Nitin Madnani. The E-rater® Automated Essay Scoring System. In *Handbook of Automated Essay Evaluation: Current Application and New Directions*, pp. 55–67. Routledge, 2013.

[19] Lawrence M. Rudner, Veronica Garcia, and Catherine Welch. An Evaluation of the IntelliMetric Essay Scoring System. *Journal of Technology, Learning, and Assessment*, Vol. 4, No. 4, pp. 1–22, 2006.

[20] Mark D. Shermis and Ben Hamner. Constrasting State-of-the-Art Automated Scoring of Essays. In *Handbook of Automated Essay Evaluation: Current Application and New Directions*, pp. 313–346. Routledge, 2013.

[21] Changhua S. Rich, M. Christina Schneider, and Juan M. D'Brot. Applications of Automated Essay Evaluation in West Virginia. In *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, pp. 99–123. Routledge, 2013.

[22] E. Mayfield and C.P.Ros. LightSIDE: Open Source Machine Learning for Text. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, pp. 124–135, 2013.

[23] Piotr F. Mitros, Vikas Paruchuri, John Rogosic, and Diana Huang. An integrated framework for the grading of freeform responses. *Proceedings of the 6th International Learning International Networks Consortium Conference*, 2013.

[24] Eun-Seo Jang, Seung-Shik Kang, Eun-Hee Noh, Myung-Hwa Kim, Kyung-Hee Sung, and Tae-Je Seong. KASS: Korean Automatic Scoring System for Short-answer Questions. *CSEDU 2014 - Proceedings of the 6th International Conference on Computer Supported Education*, Vol. 2, pp. 226–230, 2014.

[25] Min-Ah Cheon, Hyeong-Won Seo, Jae-Hoon Kim, Eun-Hee Noh, Kyung-Hee Sung, and EunYong Lim. An Automated Scoring Tool for Korean Short-Answer Questions Based on Semi-Supervised Learning. *Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pp. 59–63, 2015.

[26] 津森伸一, Alieu Dumbuya, 磯本征雄. 自由記述形式レポートの自動採点と教員評価による総合評価 (<特集>マルチメディア教材と教育・学習システム/一般). 電子情報通信学会技術研究報告. ET, 教育工学, Vol. 103, No. 135, pp. 37–42, 2003.

[27] 泉谷達庸, 片上大輔, 新田克己. 採点ルール学習とその説明機能をもつ小論文の採点支援システム. 研究報告コンピュータと教育（CE）, Vol. 2010, No. 3, pp. 1–8, 2010.

[28] 遠西学, 中村直人. e-Learning におけるレポート評価支援システムの開発. 電子情報通信学会 総合大会講演論文集 2008 年情報システム, Vol. 2008, No. 1, p. 216, 2008.

[29] 渡邊博之. ニューラルネットワークを用いた実習レポート評価支援システムの開発. 電子情報通信学会技術研究報告. ET, 教育工学, Vol. 108, No. 146, pp. 7–12, 2008.

[30] 椿本弥生, 柳沢昌義, 赤堀侃司. 出題形式や評価項目がレポート採点支援マップの可視化結果に及ぼす影響. 日本教育工学会論文誌, Vol. 33, No. 4, pp. 459–465, 2010.

[31] 村田淳哉, 片上大輔, 新田克己. SVM を利用した小論文の採点支援システム. 電子情報通信学会技術研究報告, Vol. 107, No. 428, pp. 7–12, 2008.

[32] 藤田彬, 田村直良. 文章構造解析に基づく小論文の論理性についての自動採点. 情報科学技術フォーラム講演論文集, No. 2, pp. 41–44.

[33] 勝又大介, 藤田彬, 田村直良. 文章構造解析に基づく小論文の論理構成における整然さの自動評価. 言語処理学会第 19 回年次大会発表論文集, pp. 190–193, 2013.

[34] 中島英博（編）. シリーズ大学の教授法 4 学習評価. 玉川大学出版部.

[35] 山田恒夫. MOOC と学習解析：教育革新のための情報基盤に向けて. 情報処理学会論文誌教育とコンピュータ, Vol. 1, No. 4, pp. 1–11, 2015.

[36] Piotr F. Mitros, Vikas Paruchuri, John Rogosic, and Diana Huang. An Integrated Framework for the Grading of Freeform Responses. *The Sixth Conference of MIT's Learning International Networks Consortium.*, 2013.

[37] 鈴木寛. Google ドライブのアプリおよびそのアドオンを用いた課題の作成：ルーブリックと自動採点・返却. 八戸工業大学紀要, Vol. 36, pp. 67–81, 2017.

[38] ダネル・スティーブンス, アントニア・レビ. 大学教員のためのルーブリック評価入門. 玉川大学出版部, 2014. 佐藤浩章 監訳 井上敏憲・俣野秀典 訳.

[39] 吉田武大. アメリカにおけるバリュールーブリックの動向. 教育総合研究叢書, Vol. 4, pp. 1–12, 2011.

[40] 沖裕貴, 井上史子, 林泰子. 日本の大学におけるルーブリック評価導入の方策と課題：客観的, 厳格かつ公正な成績評価を目指して. 日本教育情報学会第 28 回年会論文集, Vol. 28, pp. 166–169, 2012.

[41] 佐藤真, 香田健治. ルーブリックの開発に関するモデレーション研修の比較検討：総合的学習におけるレポート評価を通して. 教育学論究, No. 6, pp. 61–68, 2014.

[42] 松下佳代, 小野和宏, 高橋雄介. レポート評価におけるルーブリックの開発とその信頼性の検討. 大学教育学会誌, Vol. 35, No. 1, pp. 107–115, 2013.

[43] 林透, 星野晋. ルーブリック開発に関する実践的研究：初年次教育科目『山口と世界』を中心に. 大学教育, Vol. 12, pp. 10–21, 2015.

[44] Association of American Colleges and Universities. Inquiry and Analysis VALUE Rubric, 2009. https://www.aacu.org/value/rubrics/inquiry-analysis.

[45] 松本裕治, 奥村学（編）. コーパスと自然言語処理. 朝倉書店, 2017.

[46] 黒橋・河原研究室. 京都大学テキストコーパス. http://nlp.ist.i.kyoto-u.ac.jp/.

[47] 国立国語研究所. 『現代日本語書き言葉均衡コーパス』語彙表. http://pj.ninjal.ac.jp/corpus_center-/bccwj/freq-list.html.

[48] 砂川有里子. 学習辞書編集支援データベース作成について-『学習辞書科研』プロジェクトの紹介-. 日本語教育連絡会議論文集, Vol. l24, , 2012.

[49] 李在鎬. 日本語教育語彙表, 2017. http://jhlee.sakura.ne.jp/JEV.html_manual_.html.

[50] 李在鎬, 佐々木馨. 教科書コーパスを利用した難易度別コロケーション辞書の提案. 第 8 回コーパス日本語学ワークショップ予稿集, pp. 273–278, 2015.

[51] 中山浩太郎, 原隆浩, 西尾章治郎. Wikipedia マイニングによるシソーラス辞書の構築方法. 情報処理学会論文誌, Vol. 47, No. 10, pp. 2917–2928, 2006.

[52] 佐藤理史. 均衡コーパスを規範とするテキスト難易度測定. 情報処理学会論文誌, Vol. 52, No. 4, pp. 1777–1789, 2011.

[53] 佐藤理史. 日本語テキストの難易度判定ツール『帯』. *Japio YEAR BOOK 2008*, pp. 52–57, 2008.

[54] Hunter M. Breland. Word Frequency and Word Difficulty, A Comparison of Counts in Four Corpora. *Psychological Science*, Vol. 7, No. 2, pp. 96–99, 1996.

[55] 佐藤浩史, 笠原要, 金杉友子, 天野成昭. 単語親密度に基づく基本語彙の選定. 人工知能学会論文誌, Vol. 19, No. 6, pp. 502–510, 2004.

[56] 砂川有里子（編）. コーパスと日本語教育. 朝倉書店, 2016.

[57] Tomoyuki Kajiwara and Mamoru Komachi. Simple PPDB : Japanese. 言語処理学会第 23 回年次大会 発表論文集, No. C, pp. 2–5, 2017.

[58] 滝川真弘, 山名早人. 特定分野を対象とした単語重要度計算手法の提案と twitter における専門性推定への適応. 第 15 回情報科学技術フォーラム講演論文集, Vol. 15, No. 2, pp. 1–7, 2016.

[59] Stephen R. and Hugo Z. The Probabilistic Relevance Framework: BM25 and Beyond. *Journal Foundations and Trends in Information Retrieval*, pp. 333–389, 2009.

[60] 江原遥. 生コーパスからの単語難易度関連指標の予測. 言語処理学会第 23 回年次大会発表論文集, pp. 843–846, 2017.

[61] 山本恵, 梅村信夫, 河野浩之. ルーブリックに基づくレポート自動採点システム. 大学 ICT 推進協議会 2016 年度年次大会論文集, 2016.

[62] Megumi Yamamoto, Nobuo Umemura, and Hiroyuki Kawano. Automated Essay Scoring System Based on Rubric. *Studies in Computational Intelligence, Applied Computing & Information Technology*, Vol. 727, pp. 177–190, 2017.

[63] 石岡恒憲, 亀田雅之. コンピュータによる日本語小論文の自動採点システム. 電子情報通信学会技術研究報告. TL, 思考と言語, Vol. 102, No. 491, pp. 43–48, 2002.

[64] 山本恵, 梅村信夫, 河野浩之. ルーブリックに基づくレポート自動採点システムの構築. 情報処理学会第 79 回全国大会講演論文集, 2017.

[65] 山本恵, 梅村信夫. Moodle 用テキストマイニング・プラグインの試作. 私立大学情報教育協会 平成 25 年度教育改革 ICT 戦略大会, 2013.

[66] 山本恵, 梅村信夫, 河野浩之. Moodle を用いた LMS 上の自動採点システムの試作. 私立大学情報教育協会 平成 28 年度教育改革 ICT 戦略大会, 2016.

[67] 伊藤敬彦. 拡張可能なドキュメント検査ツール redpen. FIT2016(第 15 回情報科学技術フォーラム), 第 1 分冊, pp. 219–220, 2016.

[68] 浅原正幸, 加藤祥. 文書間類似度について. 自然言語処理, Vol. 23, No. 5, pp. 463–499, 2016.

[69] Christopher, D.Manning, and Hinrich Schutze. 統計的自然言語処理の基礎. 共立出版, 2017. 加藤恒昭・菊井玄一郎・林良彦・森辰則 訳.

[70] Yiming Yang. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, Vol. 1, pp. 69–90, 1999.

[71] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[72] Kimmo Kettunen. Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? *Journal of Quantitative Linguistics*, Vol. 21, pp. 223–245, 2014.

[73] 山本恵, 梅村信夫, 河野浩之. 「語彙の豊富さ」からみた学生レポートの分析・評価. *Proceedings of Moodle Moot Japan 2016 Annual Conference*, pp. 6–8, 2016.

[74] C. Udny Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 2014.

[75] 石岡恒憲, 橋本貴充, 大津起夫. 自然言語処理技術を用いたセンター試験問題の統計的解析-英語および国語の試験問題を対象として. 大学入試研究ジャーナル, No. 20, pp. 145–150, 2010.

[76] 山本恵, 梅村信夫, 河野浩之. レポート自動採点プラグインの開発と評価. *Proceedings of Moodle Moot Japan 2017 Annual Conference*, pp. 16–21, 2017.

[77] 山本恵, 梅村信夫, 河野浩之. コーパスを利用したレポート自動採点品質の向上. 情報処理学会 FIT2017 第 16 回情報科学技術フォーラム, 2017.

[78] 山本恵, 梅村信夫, 河野浩之. LMS 上の自動採点システム構築による自由記述文評価の取り組み事例. 平成 29 年度 ICT 利用による教育改善研究発表会資料集, pp. 154–157, 2017.

[79] 山本恵, 梅村信夫, 河野浩之. レポート自動採点支援用日本語語彙レベル辞書の提案-Wikipedia コーパスの利用- . 研究報告コンピュータと教育（CE）, Vol. 145, No. 12, pp. 1–7, 2018.

[80] Megumi Yamamoto, Nobuo Umemura, and Hiroyuki Kawano. Proposal of Japanese Vocabulary Words List for Automated Essay Scoring Support System Using Rubric. *The 13th International Symposium on Operations Research and its Applications*, 2018.

[81] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[82] 佐藤一誠. トピックモデルによる統計的潜在意味解析. コロナ社, 2015. 奥村学 監修.

[83] 江原遥. 単語難易度関連指標の多言語での予測. *The 31st Annual Conference of the Japanese Society for Artificial Intelligence 2017*, pp. 1–4, 2017.

[84] 岩田具治. トピックモデル. 講談社, 2016.

[85] 松河秀哉, 大山牧子, 根岸千悠, 新居佳子, 岩崎千晶, 堀田博史. トピックモデルを用いた授業評価アンケートの自由記述の分析. 日本教育工学会論文誌, Vol. 41, pp. 233–244, 2017.

[86] 芹澤翠, 小林一郎. 文書内のトピック数を考慮したトピック追跡の試み. 言語処理学会第 18 回年次大会発表論文集, pp. 1196–1199, 2012.

[87] 梅村信夫, 山本恵. LMS 向け日本語テキストマイニング支援ツール TeMP の開発と評価. 大学 ICT 推進協議会 2015 年度年次大会発表論文集, 2015.

[88] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1746–1751. Association for Computational Linguistics, 2014.

[89] Mingbo Ma, Liang Huang, Bowen Zhou, and Bing Xiang. Tree-based convolution for sentence modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2315–2325. Association for Computational Linguistics, 2015.

[90] 寺田凛太郎, 久保顕大, 柴田知秀, 黒橋禎夫, 大久保智哉. ニューラルネットワークを用いた記述式問題の自動採点. 言語処理学会第 22 回年次大会発表論文集, pp. 1196–1199, 2016.

[91] Fei Dong and Yue Zhang. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1072–1077. Association for Computational Linguistics, 2016.

付録

開発・実行環境など

表 A.1: 自動採点プラグインの開発・実行環境

LMS サーバ	データ解析サーバ	文書検査サーバ
ソフトウエア		
Ubuntu 14.04.2 LTS		
Apache HTTP Server 2.4.7		
MySQL DB Server 5.5.47	-	
PHP 5.5.9		-
Moodle 2.9.9+	R 3.2.3/ MeCab 0.996	RedPen 1.9
ハードウエア（Open VZ 仮想サーバ上のデバイス）		
CPU: Intel® Xeon® CPU E5-2640 v3 @ 2.60GHz		
3 cores	6 cores	
RAM: 2GB	RAM: 8GB	RAM: 4GB

・分類器および日本語語彙レベル辞書作成環境など

CPU： Intel® Core i7

メモリ： 16GB

OS： Ubuntu 17.0

分類器の比較テスト： Weka 3.8.1， R （version 3.4.1）

LDA 適用： Python 3.6.0， gensim 3.5.0 ライブラリーを利用

処理時間：500 トピック× 500 ワードの場合，196249sec.(54.5h)

ルーブリック関連資料

- 図 A.2 は手動採点用ルーブリックと，自動採点用ルーブリックの対応表である．
- 図 A.3 は Rubric 作成時に参照した Writing Communication Value Rubric である．AAC&U のウェブサイトよりダウンロードしたものを引用．

評価観点	達成レベルと配点					評価項目		自動採点可能項目
	0-1	2-3	4-5	6-7	8-9			
I.Content 課題の理解度と解答(記述)内容の妥当性	解答内容が，課題とは無関係である．	課題を理解し解答しているが，誤りがある．	課題を理解し解答しているが，記述が不足している．	課題を理解し的確な解答であるが，改善の余地がある．	的確な解答である。関連用語を適正に用いている。改善の必要はない．	1	論題と記述の合致度	○
						2	主要な関連語の存在	○
						3	出題意図の理解度	×
						4	内容の総合評価	×
						5	学修内容の理解度	×
Ⅱ.Structure 論理的な展開	記述内容にまとまりがない．	理論の展開に矛盾がある．	順序立てて理論を展開しているが，改善すべき点が複数ある．	順序立てて理論を展開しているが，説得力がない．	順序立てて理論を展開している．意見・主張があり，説得力がある．	6	論理性の水準	×
						7	意見・主張の妥当性	×
						8	事実と意見の区分け	×
						9	説得力	×
Ⅲ.Evidence 資料と根拠（エビデンス）の妥当性	資料を全く参照していない．根拠を示していない．	資料を参照していないが，根拠を示そうとしている．	参照しようとしている資料は相応しくない，または信頼性がない．	信頼でき，関連性のある資料を参照しているが，引用・参照の仕方に誤りがある．	当該の学問分野にふさわしく，信頼でき関連性のある資料を，うまく使いこなしている．	10	参照資料の質水準	×
						11	参照資料の関連性	×
						12	論拠資料の妥当性	×
						13	図表への説明付加	×
						14	引用量の妥当性	×
Ⅳ.Style 文章作法の遵守と適切な推敲	複数にわたってルールを守っていない．文章が全く推敲されていない．	ルールを守っていない，誤字・脱字，文体の誤りなどが複数ある．	大よそのルールを守っているが，訂正すべき点が複数ある．	訂正すべき点はないが，改善の余地がある．	よく推敲している．全く誤りがない．	15	文体の統一性	○
						16	誤字・脱字の排除	○
						17	構文の妥当性	○
						18	主述関係の妥当性	○
						19	句読点の妥当性	○
						20	冗長さの排除	○
						21	表記ゆれの排除	○
V.Skill 読みやすさ・表現の巧みさ	文章が読み辛い．明らかに文章スキルがない．	文章が長すぎるなど，複数の改善すべき点がある．	文章が概ねまとまっているが，改善すべき点がある．	文章が読みやすい．語彙が豊富である．	読み手に明確に意味を伝えることができ，読みやすい．語彙が豊富である．	22	漢字の使用率	○
						23	文長の妥当性	○
						24	語彙の豊富さ	○
						25	語彙の水準	○

図 A.2: 手動採点ルーブリックと自動採点評価項目の対応

WRITTEN COMMUNICATION VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

Definition

Written communication is the development and expression of ideas in writing. Written communication involves learning to work in many genres and styles. It can involve working with many different writing technologies, and mixing texts, data, and images. Written communication abilities develop through iterative experiences across the curriculum.

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone 4	**Milestones** 3	2	**Benchmark** 1
Context of and Purpose for Writing *Includes considerations of audience, purpose, and the circumstances surrounding the writing task(s).*	Demonstrates a thorough understanding of context, audience, and purpose that is responsive to the assigned task(s) and focuses all elements of the work.	Demonstrates adequate consideration of context, audience, and purpose and a clear focus on the assigned task(s) (e.g., the task aligns with audience, purpose, and context).	Demonstrates awareness of context, audience, purpose, and to the assigned tasks(s) (e.g., begins to show awareness of audience's perceptions and assumptions).	Demonstrates minimal attention to context, audience, purpose, and to the assigned tasks(s) (e.g., expectation of instructor or self as audience).
Content Development	Uses appropriate, relevant, and compelling content to illustrate mastery of the subject, conveying the writer's understanding, and shaping the whole work.	Uses appropriate, relevant, and compelling content to explore ideas within the context of the discipline and shape the whole work.	Uses appropriate and relevant content to develop and explore ideas through most of the work.	Uses appropriate and relevant content to develop simple ideas in some parts of the work.
Genre and Disciplinary Conventions *Formal and informal rules inherent in the expectations for writing in particular forms and/or academic fields (please see glossary).*	Demonstrates detailed attention to and successful execution of a wide range of conventions particular to a specific discipline and/or writing task (s) including organization, content, presentation, formatting, and stylistic choices	Demonstrates consistent use of important conventions particular to a specific discipline and/or writing task(s), including organization, content, presentation, and stylistic choices	Follows expectations appropriate to a specific discipline and/or writing task(s) for basic organization, content, and presentation	Attempts to use a consistent system for basic organization and presentation.
Sources and Evidence	Demonstrates skillful use of high-quality, credible, relevant sources to develop ideas that are appropriate for the discipline and genre of the writing	Demonstrates consistent use of credible, relevant sources to support ideas that are situated within the discipline and genre of the writing.	Demonstrates an attempt to use credible and/or relevant sources to support ideas that are appropriate for the discipline and genre of the writing.	Demonstrates an attempt to use sources to support ideas in the writing.
Control of Syntax and Mechanics	Uses graceful language that skillfully communicates meaning to readers with clarity and fluency, and is virtually error-free.	Uses straightforward language that generally conveys meaning to readers. The language in the portfolio has few errors.	Uses language that generally conveys meaning to readers with clarity, although writing may include some errors.	Uses language that sometimes impedes meaning because of errors in usage.

図 A.3: AAC&U の Written Communication Value Rubric

図 A.4: 教員処理実行画面の例

研究業績

論文（発表抄録を含む）

1) Moodle 用テキストマイニング・プラグインの試作，山本恵・梅村信夫，私立大学情報教育協会 平成 25 年度教育改革 ICT 戦略大会，査読無 (2013).

2) 「語彙の豊富さ」からみた学生レポートの分析・評価，山本恵・梅村信夫，Proceedings of Moodle Moot Japan 2016 Annual Conference，6-8，査読有 (2016).

3) Moodle を用いた LMS 上の自動採点システムの試作，山本恵・梅村信夫・河野浩之，私立大学情報教育協会 平成 28 年度教育改革 ICT 戦略大会，査読無 (2016).

4) ルーブリックに基づくレポート自動採点システム，山本恵・梅村信夫・河野浩之，大学 ICT 推進協議会 2016 年度年次大会発表論文集，査読無 (2016).

5) レポート自動採点プラグインの開発と評価，山本恵・梅村信夫・河野浩之，Proceedings of Moodle Moot Japan 2017 Annual Conference，16-21，査読有 (2017)..

6) ルーブリックに基づくレポート自動採点システムの構築，山本恵・梅村信夫・河野浩之, 情報処理学会 第 79 回全国大会講演論文集 (DVD)，査読無 (2017).

7) Automated Essay Scoring System Based on Rubric, Megumi Yamamoto, Nobuo Umemura and Hiroyuki Kawano, Studies in Computational Intelligence, Applied Computing & Information Technology 727, 177-190, 査読有 (2017).　※以下の口頭発表が，ジャーナルに採録されたものである.

   Automated Essay Scoring System Based on Rubric, Megumi Yamamoto, Nobuo Umemura, Hiroyuki Kawano, ACIS ACIT/BCD/CSII Conferences, ACT CITY Hamamatsu, Hamamatsu, Japan, 査読有 (2017).

8) LMS 上の自動採点システム構築による自由記述文評価の取り組み事例，山本恵・梅村信夫・河野浩之，平成 29 年度 ICT 利用による教育改善研究発表会資料集，154-157，査読無 (2017).

9) コーパスを利用したレポート自動採点品質の向上，山本恵・梅村信夫・河野浩之，情報処理学会 FIT2017 第 16 回情報科学技術フォーラム，査読無 (2017).

10) レポート自動採点支援用日本語語彙レベル辞書の提案- Wikipedia コーパスの利用；山本恵・梅村信夫・河野浩之，研究報告コンピュータと教育（CE），2018-CE-145(12)，1-7，査読無 (2018).

11) Proposal of Japanese Vocabulary Words List for Automated Essay Scoring Support System Using Rubric, Megumi Yamamoto, Nobuo Umemura, Hiroyuki Kawano, The 13th International Symposium on Operations Research and its Applications, Guiyang Confucius Hotel, Guizhou, China, 査読有 (2018).