統語的複雑性指標を用いたL2日本語学習者エッセイ評価

小畑文佳[1] 田川拓海[1] 小野雄一[2]
[1]筑波大学大学院 [2]人文社会ビジネス科学学術院人文社会科学研究群人文学学位プログラム
[3] s2320036@u.tsukuba.ac.jp tagawa.takumi.kp@u.tsukuba.ac.jp
ono.yuichi.ga@u.tsukuba.ac.jp

概要

本研究は,日本語学習者のエッセイ評価と言語的特徴量の関係性について，さらなる検討を試みたものである.従来,英語教育の分野ではL2学習者によるエッセイ評価と関連する言語特徴量が盛んに研究されている一方，日本語教育分野においては言語特徴量と評価との関連性があまり盛んに議論されていない．そこで本研究では，学習者日本語エッセイ評価に有効な言語的特徴量についてより深く検討・提案し，その評価に対する有効性を示すことを目的とした．分析では 13 の言語的特徴量を扱い，それぞれの特徴量が学習者の習熟度予測へどれほど寄与しているかを検討した.結果，一文あたりの用言数が最も有効な特徴量であることや，内容語 Guiraud 値や機能語 Guiraud 値も一定の説明力を示すことが示された．さらに，副助詞や接続助詞，格助詞の数などの機能語の数も詳細な条件として機能していることが明らかとなった.

1 はじめに

本研究では，日本語 L2 学習者エッセイにおける評価に有効な言語的特徴量を新たに提案・検討している．従来の研究では，語彙的多様性に関する言語特徴量の一部と，統語的複雑性に関する言語特徴量の一部が，それぞれ日本語エッセイにおける習熟度予測において一定の影響度を持つことを示している[1][3][6].

しかし，これらの研究で取り扱われている言語的特徴量はあくまで一部であるため，さらに新しい言語的特徴量についても検討する必要があると考えられる．今回は言語的特徴量の算出ツールを用いて，13 の言語的特徴量が習熟度予測にどれほど影響を与え得るか分析を行った.

また,分析対象となる言語的特徴量は，言語的特徴量算出ツールであるTAALED及びKWJAの機能や，日本語教育分野における文章評価の研究に基づいて選定・決定した.

2 先行研究

2．1特徴量による日本語エッセイ評価

この章では，日本語エッセイ評価において重要な言語的特徴量を検討した研究として[3][6]に加え，特徴量による評価の精度検証を行った小畑ほか[1]について紹介する.

まず[6]は，日本語教育に利用可能な指標として言語的特徴量を検討した研究の一つである．[6]では日本語の初級文法項目における生産性を数値として可視化するにあたり,機能語 Guirau 値・内容語 Guiraud 値の利用を提案している．なお生産性とは，[6]において「ある形式 X が一定の関係 R で結びつく要素の多寡の度合い P」として定義される概念である．Guiraud 値は,以下の計算式で求められる特徴量である.

$$R = \frac{Type}{\sqrt{Token}}$$

英語分野の先行研究においては生産性を評価するため,Type 頻度と Token 頻度を用いた計算方法が提案されてきた．ここで注意すべきは，同じような Type 粗頻度であっても，Token 頻度に差があるのであれば Token 頻度の上昇につれて粗頻度が上昇するため，その差を補正する必要が出てくる[6]という点である．こういった性質を踏まえ，[6]では様々な指標（Herdan の C, 共起する上位 10 項目のカバー率，標準化 TTR，修正 Perplexity，ジニ係数，エントロピー）の計算を行ったうえで，その中でも Guiraud 値が最も実用性が高いことを主張している.Guiraud 値は Token 頻度を考慮した Type 粗頻度の算出が可能であることに加え，0-30 の範囲におさまり，計算が容易である.

加えて，[6]では語彙的多様性と生産性の類似性と差異についても言及されており，生産性はある項目に接続する項目の多様性である一方，語彙的多様性は一定のテキスト内に出現した語彙の多様性を比較するものであるとしている．しかし，語彙的多様性を測定する上で重要なのは，テキストの長さ（すなわち Token 頻度）の影響を小さくしながら Type 粗頻度を比較することであり，これは生産性を算出する際に配慮しなければならない事項と共通している．このため，生産性算出のために Guiraud 値が最適であるという結論が出たことは，Guiraud 値が語彙的多様性の算出においても有効な指標である可能性が高いと考えられる．

この Guiraud 値を用いて，特徴量と習熟度の関連性について検討したのが[3]である．[3]は，習熟度に差のある作文課題を対象に言語特徴量を計測し，習熟度との関連性を報告している．分析対象としては，国際交流基金日本語国際センターで実施された訪日教師研修で用いられた，JF 日本語教育スタンダード（JFS）Bl レベルの作文課題への解答 110 件が使われた．JF 日本語教育スタンダード（JFS）は，国際交流基金（JF）によって提案された日本語学習ツールであり，低い順から A1，A2，B1，B2，C1，C2 という 6 つの言語習熟度レベルを設定している．[3]は分析にあたって，110 件の解答に対し日本語国際センターの教師 5 人で CEFR に基づく人手評価を行い，B1 の達成度に関してアノテーションを行った．結果，習熟度 B1 達成群と不達成群の間で，漢字率・文数・内容語 Guiraud 値が有意に異なることが示された．

さらに，こういった日本語エッセイにおける言語的特徴量評価研究を踏まえ，[1]は，日本語学習者による日本語エッセイの特徴量評価の精度を検証した．これは，[2]による GPT と言語的特徴量を併用した英語エッセイ評価モデルの精度検証を比較対象および参考としている．

分析対象としては，ICNALE (International Corpus Network of Asian Learners of English) と I-JAS（International Corpus of Japanese as a Second Language）を使用している．ICNALE は，神戸大学の石川らによって収集された，様々な習熟度の英語学習者によるエッセイを 1 万程度収集したコーパスである．I-JAS は国立国語研究所がデータ収集を行ったコーパスであり，日本語学習者 1,000 人による日本語の文章と発話を分析したものである．それぞれ ICNALE は日本語母語話者による英語エッセイを使用し，I-JAS は中国人母語話者による日本語エッセイを使用した．採点基準は，日本語の自動採点ツールである GoodWriting（JSPS 科学研究費補助金基盤研究 B 研究課題番号: 26284074, https://goodwriting.jp/wp/）の作文評価基準を参考に設定された．

結果としては，GPT スコアと習熟度スコアの相関は中程度に留まっている一方，Guiraud 値を中心とした特徴量によるエッセイ評価モデルは学習者習熟度との相関が強くみられたことから，言語的特徴量によるモデルの有効性を確認する結果となった．また，言語特徴量を固定効果に置いた一般化線形混合モデル・決定木分析においても同様の結果が得られた．

2．2 特徴量評価の応用

さらに，[5]は，このような特徴量評価の応用例として，言語特徴量評価を利用した，日本語短文エッセイを対象に論理的一貫性のスコアリングを行うモデルの開発を提案している．この研究では，手動で追加した言語特徴量を用いるモデル，ニューラルネットワークを用いるモデル，およびこれらを統合したハイブリッドモデルの 3 種類のアプローチを比較・評価している．

使用データとしては，人手による論理的一貫性の 5 段階評価がアノテーションされた日本語エッセイデータセットが用いられた．この評価基準には，論理的な説明がなされているか，適切な証拠が提供されているかといった要素が含まれる．これに基づいて，モデルの出力と人間評価者のスコアとの一致度を評価することで，モデルの性能を検証している．

[5]ではまず，手動特徴量を用いた回帰モデルを構築している．ここでは，文長の中央値，最大文長，句の長さ，漢字使用率，語彙的多様性（Guiraud 値），受動態の割合，単語頻度情報（BoW: Bag of Words）などの特徴量が予測に使用された．このモデルでは，ランダムフォレスト回帰を用いて論理的一貫性のスコアリングを行い，ある程度の精度を達成している．さらに，ニューラルネットワークを用いたモデルでは，日本語の事前学習済み BERT モデル（cl-tohoku/bert-base-japanese-whole-word-masking）を採用している．このモデルでは，BERT の CLS トークン（文全体を表現するトークン）の出力を基に線形層を組み合わせた回帰モデルを構築し，論理的一貫性を予測している．このアプローチにより，言語特徴量モデルと同等の精度を達成することが確認された．最終的に，これらの手法を統合したハイブリッドモ

デル（Hybrid-BERT）を提案している．このモデルでは，BERT の CLS トークン出力と手動特徴量を統合し，回帰モデルを構築することで，それぞれの手法の強みを活かしている．この結果，ハイブリッドモデルは最も高い性能を示し，Quadratic Weighted Kappa（QWK）スコアが 0.647 に達した．この値は，言語特徴量モデル（QWK=0.603）およびニューラルネットワークモデル（QWK=0.596）のスコアを上回っており，手動特徴量の統合が予測精度向上に寄与することが示された．

3 分析

今回分析対象とした言語的特徴量は以下の 13 個である．語彙的多様性に関わる特徴量は TAALED（Tool for the Automatic Analysis of Lexical Diversity）[4]を用いて抽出し，統語的複雑性に関わる特徴量は日本語解析器KWJA（Kyoto-Waseda Japanese Analyzer）の出力から求めた．また，GPT による習熟度予測は[1]と同様のプロンプトを利用して算出したスコアに基づいている．

語彙的多様性に関わる特徴量:

- 機能語 Guiraud 値(GuiraudFunction)
- 内容語 Guiraud 値(GuiraudContent)
- 漢字使用率(KanjiRate)

統語的複雑性に関わる特徴量:

- 用言数(Yougen)
- 体言数(Taigen)
- １文中の用言数(Yougen_rate)
- １文中の体言数(Taigen_rate)
- 文数(Sentence)
- 構造の最深（Max_depth）
- 副助詞の数（Fuku_jyo）
- 格助詞の数（kaku_jyo）
- 接続助詞の数（setsuzoku_jyo）

比較用

- GPT 得点（GPTScore）

データとしては，国立国語研究所がデータ収集を行ったコーパスである I-JAS（International Corpus of Japanese as a Second Language）収録の日本語学習者エッセイを使用した．I-JAS は国立国語研究所がデータ収集を行ったコーパスであり，日本語学習者 1,000 人による日本語のエッセイと発話を調査したものである．本稿では中国語母語話者によるデータを扱った．習熟度情報としては，J-CAT (Japanese Computerized Adaptive Test）の（聴解分野を除く）スコアの合計を参照した．J-CAT は，一般社団法人日本語教育支援協会が運営する，非日本語母語話者に向けた日本語能力試験であり，I-JAS のデータに学習者のスコア情報が注釈されている．

また，分析手法としてはランダムフォレストモデルおよび決定木分析を使用した．

3．1 ランダムフォレストモデル

下図は，ランダムフォレストモデルによる特徴量の重要度である．

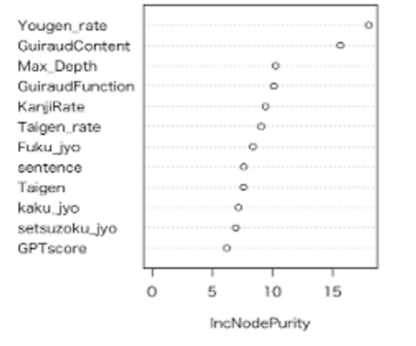

図 1　ランダムフォレストによるジニ係数

特徴量	重要度
格助詞の数	11.081151
接続助詞の数	12.494358
副助詞の数	11.715422
構造の最深	9.906784
文数	8.528016
漢字使用率	11.316876
内容語 Guiraud 値	12.248141
機能語 Guiraud 値	12.683194
体言数	10.677937
１文中の体言数	11.026697
１文中の用言数	2.312574

表 1　ランダムフォレストによるジニ係数

図 1 の横軸は IncNodePurity（ノード純度の増加量）を表し，値が大きいほど，その特徴量が学習者の習熟度の予測において重要であることを示している．この図から，1 文あたりの用言数（Yougen_rate）が最も高い重要度を持つ特徴量であることが明らかとなっている．具体的には，IncNodePurity の値が他の特徴量を大きく上回っており，学習者の習熟度を予測する際の主要な要因であることを示唆している．

また，次に重要度が高い特徴量として，内容語 Guiraud 値（GuiraudContent）や構造の最深（Max_Depth）が挙げられ，これらが習熟度の分類において大きな役割を果たしていることがわかる．

一方，助詞関連の指標（例えば，副助詞の数[Fuku_jyo]や格助詞の数[Kaku_jyo]）や，文数（Sentence）は，重要度が比較的低い結果となった．また，GPT による予測スコア（GPTscore）は重要度が非常に低く，ランダムフォレストモデルにおける予測に対してほとんど寄与しないことが示された．これらの結果から，統語的特徴や語彙多様性に関わる指標が学習者の習熟度予測において重要であり，特に 1 文あたりの用言数が顕著な影響を与えることが改めて確認された．

また，下図は言語的特徴量と学習者の習熟度（Proficiency）の分布を示すヒストグラムである．特に，1 文あたりの用言数（Yougen_rate）や内容語 Guiraud 値（GuiraudContent）が，学習者間でのばらつきが大きい一方で，分布の中心が習熟度に関連していることが見て取れる．また，語彙的多様性（例えば，GuiraudFunction や GuiraudContent）に関する指標では，分布がより幅広い範囲に及んでいるため，これらが習熟度予測における重要な変数であることが示唆される．

さらに，統語的な特徴量（例えば，体言や助詞率）については，分布が比較的集中していることに加え，これらの特徴量が分類において中程度の影響を与えることを示している．

3．2 決定木分析

図 2 は，決定木分析による特徴量の重要度と分類プロセスを視覚的に示したものである．この決定木は，学習者の習熟度（Proficiency）を予測する際に，どの言語的特徴量が分類において重要な役割を果たすかを表している．最上位ノードには，1 文あたりの用言数（Yougen_rate）が配置されており，これは全体の分類において最も重要な特徴量であることを示している．次に，左右の分岐では，それぞれ体言数（Taigen）や内容語 Guiraud 値（GuiraudContent）が重要な分岐条件として使用されている．

さらに，副助詞の数（Fuku_jyo）が 1 文あたりの体言の数が一定以上の場合に影響を与える条件として用いられている．また，機能語 Guiraud 値（GuiraudFunction）は内容語の多様性に関連する条件として使用されている．このほか，接続助詞の数（setsuzoku_jyo）や格助詞の数（kaku_jyo）も，より詳細な分岐条件として分類に貢献している．

この決定木分析の結果から，1 文あたりの用言数が分類において中心的な役割を果たす一方で，内容語 Guiraud 値（GuiraudContent）や体言，助詞関連の指標が補助的に機能していることが明らかとなった．

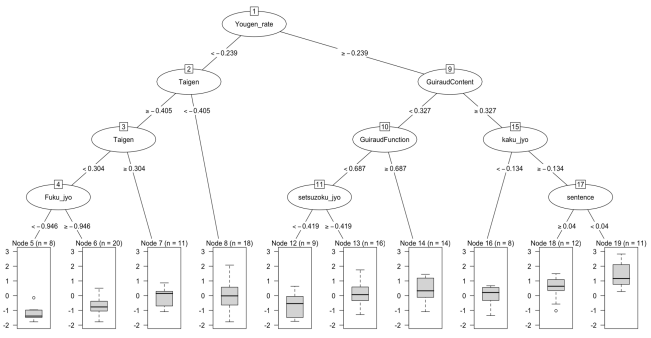

図 2　決定木

4 おわりに

本研究では，日本語学習者によるエッセイ習熟度評価を目的として，評価に有効な言語特徴量のエッセイ評価モデルにおける有効性を検討した．まず分析 1 における有効な言語特徴量の検討では，一文あたりの用言数（Yougen_rate）が最も有効な特徴量であることが示された．この特徴量は，学習者の習熟度（Proficiency）との間に有意な正の相関を示しており，統語的複雑性を捉える指標として有用であることが明らかになった．また，内容語 Guiraud 値(GuiraudContent)や機能語 Guiraud 値（GuiraudFunction）も一定の説明力を示したが，一文あたりの用言数ほどの寄与には及ばなかった．また，GPT によるスコアは習熟度予測に際しほとんど寄与しなかった．

今後の展望としては，まず他言語話者および他言語母語話者によるデータセットにおける言語特徴量による評価の有効性については検証の必要があると考えられる．特に特に漢字文化圏と非漢字文化圏では，漢字の使用率といった特徴量について習熟度との関連性が変化する可能性がある．また，[5]のように特徴量評価の応用として構築する評価システムについても検討していきたい．

謝辞

本研究は JSPS 科研費 21H00905 の助成を受けたものです．

参考文献

1．**Ayaka Obata**, **Takumi Tagawa**, and **Yuichi Ono**. : Assessment of ChatGPT's validity in scoring essays by foreign language learners of Japanese and English. Proceedings of 15th International Congress on Advanced Applied Informatics, 2023 December 11-13, Bali, Indonesia.

2．**Atsushi Mizumoto** and **Masaki Eguchi**. : Exploring the potential of using an AI language model for automated essay scoring．Res．Methods Appl. Linguist., vol. 2，no. 2，pp. 100050, 2023.

3．**Hiroaki Hatano.** : An investigation of the linguistic features of B1 level of the JF Standard for Japanese Language Education: An analysis based on the writing data by Japanese language learners. Learner Corpus Stud. Asia World, vol. 3，pp. 189-206，Mar. 12, 2018.

4．**Kyle, K., Crossley, S. A., & Jarvis, S.** : Assessing the validity of lexical diversity using direct judgements．Language Assessment Quarterly 18(2)., pp. 154-170, 2021.

5．**Nakamoto, S．& Shimada, K.** : Automated scoring of logical consistency of Japanese essay. Proceedings of the 24th International Conference on Artificial Intelligence in Education.

6．**Naoki Nakamata**. : Visualization of the productivity of function words in basic Japanese grammar syllabus. The Mathematical Linguistic Society of Japan., vol. 2, no. 8, pp. 275-295, 2015.