英語 L2 ライティングにおける LLM 校正による文法エラーの残存

小野雄一
筑波大学　人文社会系
ono.yuichi.ga@u.tsukuba.ac.jp

概要

本研究は，生成 AI による校正(revision)が L2 学習者テキストの文法エラー分布をどの程度中和するかを，LanguageTool のカテゴリ別修正率(fix rate)で定量化する. ICNALE の JPN/CHN 各 800 編(計 1600 編)に対し, gpt-4o-mini (temperature=0)で意味保持を課した strong 校正を適用し，校正前後のエラーカテゴリを比較した．結果，GRAMMAR，PUNCTUATION はほぼ天井値の修正率を示す一方，REDUNDANCY，COLLOCATIONS, REPETITIONS_STYLE, TYPOS は修正率が相対的に低く残存した．母語・熟達度の効果は限定的であり，修正の成否は学習者属性より誤りタイプに強く依存することが示唆された．

1 はじめに

同一の言語で書かれたテキストであっても，書き手の母語に由来する特徴的なパターンが残存することは広く知られている．こうした L1 fingerprint は，語彙選択，冗長性，談話構造，機能語の分布などに現れ，法言語学や著者同定の分野を中心に研究が蓄積されてきた[1,2]. L2 ライティング研究においても，Native Language Identification (NLI)の枠組みを通じて，学習者の母語に由来する書き言葉の特徴が検討されてきた．先行研究[3]は，母語影響が文法誤りにとどまらず，語彙的選択や情報配置といった高次の言語的選好にも及ぶことを報告している．近年，生成 AI (LLM)による自動修正が L2 ライティングで広く用いられているが，こうした修正が L1 fingerprint をどの程度中和するのかは十分に解明されていない．特に，規則ベースの文法的誤りと，語彙選択や冗長性といった選択性の高い誤りが同様に処理されるのかについては明らかになっていない．本研究はこの点に着目し，LanguageTool [4]によって検出される文法エラーカテゴリを対象として，生成 AI による強力な校正(strong revision)の効果をカテゴリ別修正率(fix rate)に基づいて検討する．

2 先行研究

2.1 L1 fingerprint と Native Language Identification (NLI)

第二言語習得研究では，学習者の産出言語において，母語に由来する体系的な特徴が残存することが古くから指摘されてきた[5]．このような母語影響は，個別の誤用にとどまらず, 語彙選択, 機能語の分布, 統語的選好といった複数の言語レベルにわたって観察されることが報告されている[6].

こうした母語由来の特徴を計算的に捉える枠組みとして発展してきたのが，Native Language Identification (NLI)研究である．NLI 研究では，学習者テキストから抽出された語彙 n-gram，機能語，構文的特徴などを用いて，書き手の母語を高精度で推定できることが示されてきた[1, 7]．これらの研究は，学習者の母語影響が偶発的な誤りの集合ではなく，統計的に安定した fingerprint としてテキスト全体に分布していることを明らかにしている．

特に重要なのは，NLI 研究の多くが，誤り訂正や正規化を施した後のテキストにおいても，母語推定が一定程度可能であることを報告している点である[2, 3]．このことは，表層的な文法誤りが修正された後にも，母語に由来する言語的特徴が完全には消失しない可能性を示唆している．

2.2 自動誤り検出・文法エラー分類と L2 ライティング研究

L2 ライティング研究においては, 学習者の誤りを体系的に捉えるために，自動誤り検出器やエラー分類体系が広く用いられてきた．従来の研究では，人工的に付与された誤りタグや人手アノテーションに基づく分析が主流であったが，近年では LanguageTool や Grammarly に代表される自動文法検出器が，誤りの大規模分析に用いられるようになっている[8,9].

特に LanguageTool については，GRAMMAR，PUNCTUATION，TYPOGRAPHY，REDUNDANCY

などのカテゴリに基づき，誤りを比較的一貫した基準で検出する点に特徴がある．そのため，学習者間・母語間・熟達度間の比較を行う上で，再現性の高い測定手段として利用可能である．

一方で，自動誤り検出器が捉える「誤り」は，必ずしもすべてが規則ベースの文法逸脱に限定されるわけではない．冗長性や反復，語彙選択の不自然さといったカテゴリは，文法性と文体性の境界に位置しており，学習者の母語的選好や書き方の癖を反映する可能性がある．この点において，自動誤り分類は，L1 fingerprint の残存領域を間接的に捉える手がかりを提供すると考えられる．

2.3 生成 AI による校正と文法エラー修正の限界

近年，生成 AI，特に大規模言語モデル(LLM)の発展により，L2 ライティングにおける自動校正・リビジョンが急速に普及している．先行研究は，LLM による校正が，文法誤りの削減や表現の自然化において高い性能を示すことを報告している．

しかし，これらの研究の多くは，修正後テキストの全体的な品質評価や学習者の態度調査に焦点を当てており，どの種類の文法エラーがどの程度修正されるのか，またすべての誤りカテゴリが同様に正規化されるのかという点については，十分に検討されていない．特に，生成 AI による強力な校正が，規則ベースの文法誤りと，語彙選択や冗長性に関わる誤りを同一の精度で処理するのかどうかは，未解明の課題である．この点を明らかにすることは，生成 AI が「文法校正器」としてどこまで機能しうるのか，また L1 fingerprint がどのレベルで残存するのかを理解する上で重要である．

2.4 本研究の位置づけ

以上の先行研究を踏まえると，生成 AI による文法校正に関して，以下の点が十分に明らかにされていない．

(1) 自動誤り検出器が定義する誤りカテゴリは，LLM による強力な校正によって一様に修正されるのか
(2) 修正に対して頑健な誤りカテゴリは存在するのか
(3) そのようなパターンは，学習者の母語や熟達度によって異なるのか

本研究は，LanguageTool によって検出された文法エラーに注目し，カテゴリ別修正率(fix rate)を用いて，生成 AI による強力な校正が L1 に特有の誤り分布をどの程度中和するのかを定量的に検討する．

3 方法

3.1 コーパスと対象データ

本研究では，アジア圏英語学習者のエッセイを収録した ICNALE(International Corpus Network of Asian Learners of English) [10]を使用した．分析対象は，日本語母語話者(JPN, 800 エッセイ)および中国語母語話者(CHN, 800 エッセイ)の学習者エッセイであり，CEFR レベルに基づき A2，B1，B2 の 3 段階に分類されたデータを用いた．付録の表 1 に記す。

各エッセイは，同一トピックに基づく意見文として収集されており，内容的制約が比較的一定に保たれている．この点は，生成 AI による修正効果を，内容差ではなく言語的側面の変化として比較する上で適している．

学習者の熟達度は CEFR に基づき A2，B1，B2 の 3 段階に分類されている．本研究では，母語(JPN/CHN)および熟達度(A2/B1/B2)を主要な説明変数とし，生成 AI による校正が文法エラー分布に与える影響を比較する．

3.2 生成 AI による校正条件

本研究では，生成 AI による校正の効果を明確に検証するため，ミニマムな修正ではなく strong revision を課した．strong revision では，原文の意味を厳密に保持した上で，流暢で自然な英語に「全面的」に書き直すことを生成 AI に指示した．

Strong revision: *Rewrite into fluent academic English while strictly preserving meaning.*

校正には OpenAI の Chat Completions API を用い，各エッセイに対して同一の system/user プロンプトを適用した．生成条件は以下の通りである．

- モデル：gpt-4o-mini
- temperature：0
- system メッセージ：役割・制約を固定
- user メッセージ：意味保持を明示的に指示

生成 AI による出力は確率的生成に基づくものであり，理論上は完全な決定性は保証されない．ただし，本研究では temperature を 0 に設定し，プロンプトおよび生成条件をすべて固定することで，処理の再現性と条件間比較の妥当性を確保した．

3.3 文法エラーの検出とカテゴリ化 (LanguageTool)

生成 AI による校正前後のテキストに含まれる文法エラーは自動文法検出器 LanguageTool を用いて検出した．LanguageTool は，ルールベースおよび統計的手法を組み合わせたオープンソースの文法チェックツールで，再現性の高い誤り検出が可能である．

本研究では，LanguageTool が出力するエラーカテゴリのうち，以下の主要カテゴリを分析対象とした．

- GRAMMAR
- PUNCTUATION
- TYPOGRAPHY
- CASING
- REDUNDANCY
- REPETITIONS_STYLE
- COLLOCATIONS
- TYPOS

これらのカテゴリは，規則ベースの文法逸脱から，語彙選択や冗長性に関わるスタイル的逸脱までを含んでおり，生成 AI による校正がどのレベルの誤りに介入しているかを比較するのに適している．

3.4 修正率(fix rate)の定義

修正率(fix rate)は，修正前に検出されたエラー数(#errors_pre)と，strong 校正後に残存したエラー数(#errors_strong)に基づき，次式(1)で定義した．

$$\textbf{fix rate} = 1 - \textbf{(\#errors\_strong / \#errors\_pre)} \quad \ldots (1)$$

この式において，値が 1 に近いほど，当該カテゴリの誤りがほぼ完全に修正されたことを意味する．なお，修正前にエラーが検出されなかったカテゴリについては，fix rate を算出せず，分析から除外した．

3.5 統計分析

カテゴリ別 fix rate を応答変数とし，母語(JPN/CHN)，熟達度(A2/B1/B2)，およびその交互作用を固定効果とした線形モデルを構築した．分析はカテゴリごとに実施し，生成 AI による校正効果が，誤りタイプによってどの程度異なるかを検討した．

本研究の分析は，文法エラーの検出および修正という限定された側面に焦点を当てることで，生成 AI が「文法校正器」としてどの程度機能するのか，またどの誤りカテゴリが修正に対して頑健であるのかを明確にすることを目的としている．

4 結果

本節では，LanguageTool により検出された文法関連エラーを対象として，ChatGPT による strong 校正が学習者エッセイに及ぼす影響を検討する．特に，(1) エラーカテゴリ別の修正率(fix rate)の全体傾向，(2) 母語(L1)および熟達度による差異，(3) strong 校正後に残存するエラーカテゴリに焦点を当てる．

4.1 カテゴリ別修正率(fix rate)の全体傾向

図 1 は，全 1600 編(JPN/CHN 各 800 編)を対象に，strong 校正による文字レベルの修正量(diff_char_strong)と LanguageTool に基づく修正率(fix rate)との関係を，母語(JPN/CHN)および熟達度(A2/B1/B2)別に示したものである．各点は文書単位の観測値を表し，回帰直線と 95%信頼区間を併せて示している．全体として，修正量が増加しても修正率は高水準で維持される傾向が確認されたが，その関係

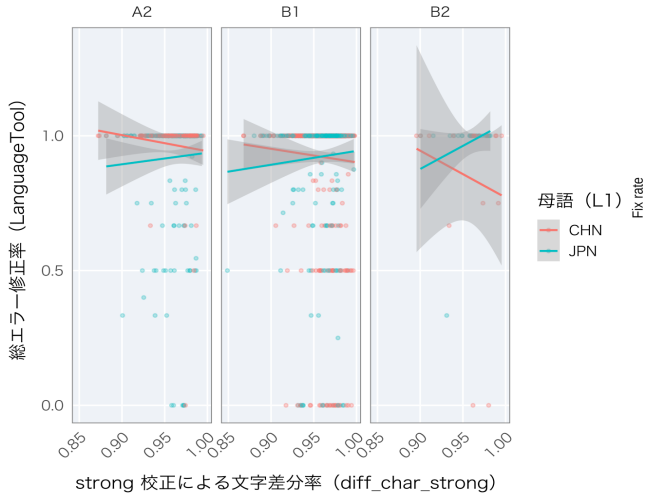

図 1　strong 校正の改変量と LanguageTool 修正率

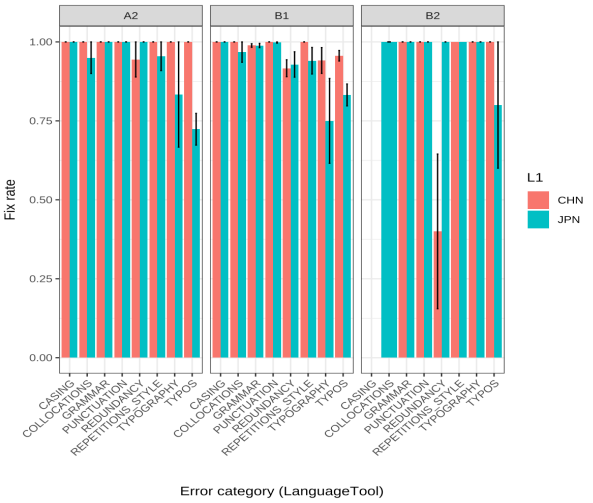

図 2　母語(L1)・熟達度・カテゴリ別修正率

は熟達度によって異なっていた．特に B2 レベルでは母語間で回帰直線の傾きに差がみられ，strong 校正の効果が一様ではない可能性が示唆される．一方，A2 および B1 レベルでは，修正量と修正率の関係は比較的安定しており，母語による大きな乖離は認められなかった．

図 2 は，カテゴリ別修正率の分布を詳細に確認するため，改変率が高かった文書(pre→strong の文書類似度が低いもの)を各セル(L1×熟達度)から 20 編ずつ抽出した計 120 編を対象として，LanguageTool のエラーカテゴリ別修正率を示したものである．修正前に当該カテゴリのエラーが検出されなかった文書は分析から除外した．その結果，GRAMMAR および PUNCTUATION ではすべての条件で修正率がほぼ天井値(約 0.98－1.00)に達しており，strong 校正が基礎的な文法・形式エラーに対して極めて高い修正能力を有することが示された．CASING や TYPOGRAPHY でも高い修正率が観察されたが，検出エラー数が少ない条件を含むため解釈には注意を要する．

さらに，カテゴリ別修正率を従属変数とし，母語，熟達度，およびその交互作用を固定効果とする線形モデルを構築した結果，いずれの効果も有意ではなかった(すべて $p > .30$)．このことは，strong 校正が母語や熟達度に依存せず，全体として一様に適用されていることを示している．

4.2 修正率が低く留まるエラーカテゴリ

一方，REDUNDANCY，REPETITIONS_STYLE，COLLOCATIONS，TYPOS では，strong 校正後もエラーが一定数残存し，修正率は概ね 0.7－0.9 であった．特に REDUNDANCY および COLLOCATIONS では，母語や熟達度条件によるばらつきが相対的に大きかった．これらのカテゴリは，明確な規則違反というよりも語彙選択や情報配置といった選択的要素を含むため，完全な正規化が行われにくい可能性が示唆される．また TYPO では，条件によって修正率が 0.7 を下回る例も見られ，LLM による書き換えが必ずしもすべての表層的誤記を一貫して除去するわけではないことが示された．

4.3 母語(L1)および熟達度の効果

カテゴリ別修正率を従属変数とした分析の結果，GRAMMAR や PUNCTUATION では，母語・熟達度・交互作用のいずれについても有意な効果は認められなかった．これは，strong 校正がこれらのカテゴリに対して一律に適用されていることを示している．

一方，REDUNDANCY や COLLOCATIONS では条件間のばらつきが相対的に大きく，母語や熟達度による差異が視覚的に確認された．ただし，これらの差は限定的であり，全体的な正規化傾向を大きく左右するものではなかった．

4.4 strong 校正後に残存するエラーの特徴

strong 校正後にも検出されたエラーは，主として冗長性，反復，コロケーション選択，軽微な誤記に集中していた．これらは明確な規則違反ではなく，複数の許容可能な選択肢の中からの選択に関わるものである．母語間の差は全体として小さいものの，冗長性や語彙選択に関わる誤りでは，母語に由来する選好が strong 校正後も部分的に残存した．

以上から，strong 校正は規則ベースの文法誤りをほぼ完全に正規化する一方で選択性の高い言語的特徴は完全な中和に至らないことが示された．

5. 考察

本研究は，GRAMMAR や PUNCTUATION といった規則ベースの文法誤りは，母語や熟達度にかかわらずほぼ完全に修正される一方で，REDUNDANCY や COLLOCATIONS など語彙的・文体的選択を伴う誤りカテゴリは，強力な校正後も一定程度残存することが明らかとなった．B2 レベルでは，文法的逸脱そのものは相対的に少なく，LanguageTool が検出可能なエラーは限定的である一方，生成 AI は文全体の情報配置やスタンス表現を最適化する方向で改変を行う．その結果，改変量は大きくなるものの，エラー削減率としては必ずしも増加しない状況が生じうる．

これらの結果は，生成 AI が高性能な文法校正器として機能する一方で，L1 fingerprint を全面的に消去する存在ではないことを示している．英語教育学的には生成 AI を文法的正確性の担保装置として活用しつつ，語彙選択や冗長性といった領域については学習者自身の気づきや指導的介入が依然として重要であることを示唆している．今後は，第二言語習得論的に習得困難とされる文法項目の修正可能性・妥当性や，談話構造やスタンス表現など，文法エラーを超えた言語レベルに分析を拡張することで，生成 AI と母語影響の関係をより包括的に明らかにする必要がある．

謝辞

本研究は JSPS 科研費 24K00081 の助成を受けたものである．

参考文献

[1] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Automatically Determining an Anonymous Author's Native Language. Intelligence and Security Informatics, **Lecture Notes in Computer Science**, Vol. 3495, pp. 209–217, Springer, 2005.

[2] Julian Brooke and Graeme Hirst. Robust, Lexicalized Native Language Identification. **Proceedings of COLING 2012**, pp. 391–408, Mumbai, India, 2012.

[3] Shervin Malmasi and Mark Dras. Large-Scale Native Language Identification with Cross-Corpus Evaluation. **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1403‑1409, 2015.

[4] Marcin Miłkowski. Developing an Open-Source, Rule-Based Proofreading Tool. **Software: Practice and Experience**, Vol. 40, No. 7, pp. 543–566, 2010.

[5] Terence Odlin. Language Transfer. **Cambridge Applied Linguistics Series**, Cambridge University Press, 1989.

[6] Scott Jarvis and Aneta Pavlenko. **Crosslinguistic Influence in Language and Cognition**. Routledge, 2008.

[7] Joel Tetreault, Daniel Blanchard, and Aoife Cahill. A Report on the First Native Language Identification Shared Task. **Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications,** pp. 48–57, 2013.

[8] Dana R. Ferris. **Treatment of Error in Second Language Student Writing**. University of Michigan Press, 2002.

[9] Jim Ranalli. Automated Written Corrective Feedback: How Well Can Students Make Use of It? **Computer Assisted Language Learning**, Vol. 31, pp. 1–22, 2018.

[10] Shin'ichiro Ishikawa. **The ICNALE Guide: An Introduction to Learner Corpus Study on Asian Learners' L2 English**. Routledge, 2023.

付録

表１　分析対象のテキスト

	A2	B1_1	B1_2	B2+	合計
CHN	50	232	105	13	400
JPN	154	179	49	18	400
合計	204	411	154	31	800