日本語学習者作文の一貫性評価における日本語の誤りの影響

甲斐 宥太[1]　横野 光[1]
[1] 明星大学
21j5032@stu.meisei-u.ac.jp　hikaru.yokono@meisei-u.ac.jp

概要

本研究では，日本語学習者の作文の一貫性評価に対して，作文中に含まれる日本語の誤りがどの程度影響を与えるかについての分析を行った．日本語学習者縦断作文コーパスに付与されている印象評価データの一貫性の項目を用いて，元の作文と誤り訂正を行った作文に対する大規模言語モデルによる一貫性評価を行い，人手の評価値の平均二乗誤差を求めた．実験の結果，誤り訂正による一貫性評価の性能改善は見られず，一貫性の評価に対する誤りの影響は明確にはないということが示された．

1　はじめに

日本語学習者の数は増加傾向にあり，それに伴って日本語教師の負担も増大している．特に作文指導や評価は学習者一人一人の文章を精読し，内容や構成，表現について個別にフィードバックを行う必要があるため，時間と労力を要する作業である．このような状況において，教育現場を支援する手段として自動作文評価を活用する試みが進んでいる．

近年の大規模言語モデル (LLM) の発展により自動作文評価の精度が向上しており，先行研究では，文法的項目や文章全体の質を評価する試みが多く行われてきた．一方で，意味的なまとまりを指す一貫性 [1] という評価項目は，複数の文や段落にまたがる意味的な関連性を扱うため，自動的な評価は比較的困難である．一貫性は内容に関する項目であるため，誤字などの表層的な誤りとは切り離して評価されるべきだと考えられる．しかし，評価するためにはその作文を読む必要があり，その際に作文に含まれる誤りが評価者の一貫性の評価に影響を与える可能性がある．日本語学習者が執筆する作文には，誤字脱字や漢字の誤り，助詞や活用の誤りなどが多く見られる．これらの誤りは，学習過程において自然に生じるものであり，日本語教育において重要な分析対象である一方で，文章の内容理解や評価を困難にする要因となり得る．

そこで本研究では一貫性の評価において作文中の日本語の誤りがどの程度影響を与えるかについて，日本語学習者縦断作文コーパス W-CoLeJa[2] に付与されている人手による一貫性の評価データを用いて分析を行う．

2　関連研究

従来の作文自動評価では，語彙の多様性や文法的正確性，文長などの表層的特徴量を用いた評価が中心であったが，近年では LLM の発展により，談話レベルでの一貫性の扱う手法が注目されている．Li らは，LLM を用いて，作文教育における学生の行う作文の改訂の質を自動評価し，人手評価との中程度の一致を示している [3]．この研究では，詳細な評価基準をプロンプトに含めることで，改訂の質に関する判断が一定程度可能であることが示された一方，Chain-of-Thought の導入が必ずしも評価精度の向上につながらない点も報告されている．Naismith らは，LLM を用いて文章の一貫性を自動的に評価する手法を提案している [4]．この研究は，一貫性の評価に対して，近年の LLM が有効に機能する可能性を示している．同じく一貫性の自動評価について，説明可能性の高いモデルによる手法と LLM による手法を比較している Azrou らの研究においても，LLM はほかのモデルを大きく上回る性能を示している [5]．

一方で，これらの研究は学習者の誤用を考慮した設定で行われてはいない．誤りは形態素解析などにおいても解析誤りを引き起こすことが多く [6]，日本語学習者作文の表層的誤りが多く含まれる文章に対して，同様の手法がどの程度有効であるかについては，十分に検討されていない．

3　印象評価データ

日本語学習者縦断作文コーパス W-CoLeJa[2] は中国などの海外の大学で日本語を学習している学生の

日本語作文を 4 年間 (一部は 3 年間) にわたって収集した学習者コーパスである．作文は体験文，説明文，意見文のそれぞれについて設定されたテーマで書かれている．

この学習者作文の一部に対して，印象評価情報として，作文がよく書けているかどうかという観点の作文全体の総合評価 (10 段階) や，一貫性をはじめとして構成や正確さなどの 11 項目の個別の評価 (5 段階) と各評価に対する理由がアノテーションされている (以降，印象評価データと呼ぶ)．アノテーションはクラウドソーシングを用いて行っており，1 作文につき 6 人または 7 人の作業者が評価を行っている．作業者は作文コンテストの審査員の立場として評価するように指示されている．作業者は日本語の母語話者であり，日本語教育の経験や非母語話者の作文の添削の経験については要件としていない．

本研究ではこの印象評価データの一部の 118 件を使用している．母語・地域と学年別の内訳を表 1 に示す．

表 1　作文データの内訳

母語・地域 \学年	1 年	2 年	3 年	4 年	合計
中国語・中国	10	10	10	10	40
中国語・台湾	10	10	10	10	40
ベトナム語・ベトナム	9	9	10	10	38
合計	29	29	30	30	118

印象評価データの個別評価のうち一貫性の評価結果 (1(平均より非常に劣っている)~5(平均より非常に優れている)) を分析に使用する．作業者に与えられた一貫性の説明は “文章の筋が通っていてまとまりがある” であり，これを参考にして評価を行っている．

一貫性の判定基準は人によって異なっていると考えられる．評価の揺れを調べるために，各作文に付与されている一貫性評価の値の最大値-最小値を求めた．そのヒストグラムを図 1 に示す．作業者間で評価が一致した作文は極めて少なく，差が 3 以上のものが半数以上を占めている．特に差が 4 の作文は同じ作文でも極端に評価が分かれているものである．その要因としては，一貫性を客観的に定義することが困難であるという点と，クラウドソーシングで行ったアノテーションであるため作業者に対して詳細な指示を行うことが困難であったという点が考えられる．また，図 1 は最大値の値で色分けしており，例えば差が 3 の作文の大半は最大値が 4 であり，平均よりも優れているという評価と平均よりも非常に劣っているという反対の印象の評価が混在している．

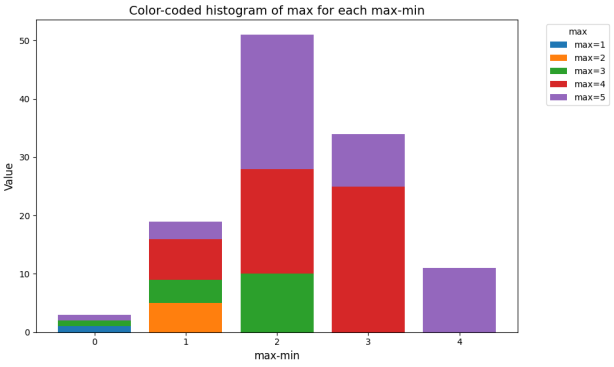

図 1　一貫性評価値の最大値-最小値のヒストグラム

印象評価データに含まれている作業者の一貫性評価の判定理由には日本語の誤りについての記述が数件見られたが，それが一貫性の評価値の要因になったという記述はなかった．このことから，評価者は一貫性の評価においては日本語の誤りの影響はほとんど受けていないと見ることができる．一方で，コメントには “筋が通っている”，“矛盾している”，“一貫している” といった記述が多く見られた．このことからは，一貫性の評価では具体的な箇所を評価の根拠とするということが行われにくいということが示唆される．

4　LLM による一貫性評価

本研究では，日本語学習者の作文に対して誤用や文法誤りなどの日本語の誤りが一貫性の評価においてどのような影響を与えるかについて，必要最低限の誤り訂正を行った作文と元の作文のそれぞれについて一貫性の自動評価を行いその結果で分析する．誤り訂正と一貫性評価にはいずれも LLM を用いる．

4.1　誤り訂正

一貫性は内容に関する評価のため，誤りの訂正は，誤字脱字，漢字，助詞，活用，時制，文体といった表層的な誤りのみを対象とし，文の順序や内容の追加・削除，意味の書き換えといった，文章全体の意味構造や一貫性そのものに影響を与えると考えられる訂正は行わない．誤り訂正には OpenAI[1)] の GPT-4o を用いた．使用したプロンプトを図 2 のプロンプトに示す．

また，このプロンプトによる誤りの訂正例を図 3 に示す．誤りの訂正を必要最低限にとどめているため全ての誤りが訂正できているわけではない．

1)　http://openai.com

次の作文から誤字脱字，漢字，時制，文体，「てにをは」の間違いのみを修正し他を変えずに本文のみを出力してください

作文

図 2　誤用修正のプロンプト

訂正前　この手帳を書きたのですが，待ちきれなかった．
訂正後　この手帳を書きたくて，待ちきれなかった．

図 3　誤りの訂正例 (ID:CCC045-A1-02-04 より引用)

4.2　一貫性評価

学習者の作文を入力とし，LLM による一貫性の評価を行い，印象評価データに付与されているものと同様の評価値を出力させる．プロンプトには指示の他に，一貫性に関する知識として，一貫性の説明，印象評価情報としてアノテーションされている一貫性評価の判断理由から抽出した評価観点，高等学校の国語の学習指導要領 [7] における一貫性に関する記述[2)]を与える．また，few-shot の事例として評価値の平均が 4 と 2 に近く，かつ作業者間の揺れが小さい作文から作文の長さが近いものを 2 件選び，プロンプトに加えている．

使用したプロンプトのひな形を図 4 に示す．

指示
あなたは日本語学校の先生です．以下に示す説明に従って，文章の一貫性を評価してください．1〜5 の 5 段階で数値のみ返してください．

一貫性の説明

判断理由中の評価観点

指導要領における記述

few-shot の事例

作文

図 4　一貫性評価のプロンプトのひな形

2)　各項目の具体的な内容は付録 A を参照．

5　分析

5.1　実験設定

一貫性評価モデルのプロンプトに与える few-shot の事例とした 2 件を除いた 116 件の作文を用いて，元の作文と誤り訂正を行った作文のそれぞれに対して LLM による一貫性評価を行った．印象評価データの一貫性の評価値を正解データとするが，その評価値は 1 作文につき 6，7 件付与されているため，その平均値を正解としている．評価尺度は予測値と正解との平均二乗誤差を用いた．

図 1 に示したように一貫性の評価値は作業者による揺れが比較的多く，この揺れの大きさが一貫性評価にどのような影響を与えるかを分析するため，作文に付与されている評価値の最大値-最小値 (max-min) が 3 以上の作文の集合と 3 未満の作文の集合に分割し，それぞれで評価を行った．

一貫性評価のプロンプトには指示以外に 3 つの一貫性に関する知識を加えている．使用する知識の組み合わせとして以下の 5 種類の設定を用いた．また，それぞれの設定に対して zero-shot の場合と few-shot の場合での結果を生成している．

設定 1　知識不使用 (指示のみ)
設定 2　一貫性の説明，評価観点，指導要領
設定 3　一貫性の説明，評価観点
設定 4　一貫性の説明，指導要領
設定 5　一貫性の説明

5.2　結果と考察

実験結果を表 2 に示す．

表 2　実験結果 (太字は最良を示す)

		max-min >=3		max-min <3	
		訂正前	訂正後	訂正前	訂正後
設定 1	zero-shot	0.7356	0.8837	0.8429	1.2192
	few-shot	0.7218	0.8721	0.8684	0.8530
設定 2	zero-shot	0.9377	1.0118	0.9274	0.9013
	few-shot	0.7525	0.9567	0.8577	0.9261
設定 3	zero-shot	0.9768	1.0488	0.8034	1.1360
	few-shot	1.2340	1.0149	**0.6659**	1.0106
設定 4	zero-shot	0.6890	0.7874	0.9422	0.9637
	few-shot	0.8636	1.1620	0.8396	0.9663
設定 5	zero-shot	0.8477	1.0710	0.9851	0.9556
	few-shot	**0.6520**	0.8498	0.8013	1.0743
平均		0.8411	0.9658	0.8534	1.0006

まず，本研究の提案手法の中心である誤用修正の効果について検討する．平均二乗誤差の平均を見る

と，max-min>=3，max-min<3 のいずれにおいても訂正前の方が訂正後よりも小さくなっており，平均的には誤り訂正による一貫性の評価性能の向上は見られなかった．

誤り訂正が一貫性そのものに影響を与えないように訂正内容を限定的なものにとどめており，訂正前後の作文 118 組に対して文章の意味的類似度として求めた Sentence-BERT[3] によるコサイン類似度の平均は 0.9884 であった．この結果から，訂正前後による内容の意味的な変化は抑えられていたことが確認できる．また，表記上の差異を測るために編集距離を算出し，訂正前の作文の文字数で割った文字数当たり編集距離の全体平均は 0.05660 であった．このことからは実際は誤り訂正は作文の内容を大きく変化させない範囲で行われていたと考えられる．

表 3 に zero-shot と few-shot での平均を示す．max-

表 3 zero-shot, few-shot での平均

	max-min >=3		max-min <3	
	訂正前	訂正後	訂正前	訂正後
zero-shot	0.8374	0.9605	0.9002	1.0352
few-shot	0.8448	0.9711	0.8066	0.9661

min >=3 ではやや zero-shot の方が良く，max-min<3 では few-shot の方が良い結果となっている．しかし，各設定での比較では一貫してどちらかが良いという結果にはなっておらず，few-shot は効果的ではあるが，安定しているとは言えないということが示された．

ユーザープロンプトの構成要素に着目すると，最も誤差が小さかったのは max-min>=3 では設定 5，max-min<3 では設定 3 であった．いずれも全ての知識を使用した設定ではなく，必ずしも多くの説明や評価基準を与えたプロンプトほど精度が高くなるわけではないことが分かる．また，いずれも指導要領を使用していないものである．今回使用した指導要領は高校の国語のものであり，記述されている一貫性は主に論述文に対するものである．これは今回の対象である日本語学習者の作文とは性質が異なるものであり，この実験の結果からは一貫性については文章の種類などによって考慮すべき要素が異なっているということが示唆される．

6 おわりに

本研究では，日本語学習者の作文を対象にした一貫性評価に対して，作文に含まれる日本語の誤りが評価に影響を与えるかどうかについての分析を行った．人手によって付与された一貫性の評価値を用いて行った元の作文と誤り訂正を行った作文のそれぞれに対する LLM による一貫性評価では，誤り訂正による性能改善は見られなかった．しかし，誤り訂正に関しては，正しく訂正されたかといった誤り訂正の性能評価は行っておらず，この点についても分析を行う必要がある．

また，今回使用した印象評価データでは作業者に与えられた指示は簡潔なものであり，また，作業者も日本語教育の専門家ではなかった．この点に関しても，日本語の作文教育において重視される一貫性とは何かなどに関して，現場では一貫性に対してどのような指導を行っているのかといった調査などから，より詳細な定義を考える必要がある．

3） https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2 を使用

謝辞

本研究は JSPS 科研費 21H04417 の助成を受けたものです．

参考文献

[1] 庵功雄. 日本語におけるテキストの結束性の研究. くろしお出版, 2007.

[2] 石黒圭, 烏日哲（編）. 学習者コーパスの設計と構築. 研究社, 2025.

[3] Tianwen Li, Zhexiong Liu, Lindsay Matsumura, Elaine Wang, Diane Litman, and Richard Correnti. Using large language models to assess young students' writing revisions. In Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, editors, **Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)**, pp. 365–380, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[4] Ben Naismith, Phoebe Mulcaire, and Jill Burstein. Automated evaluation of written discourse coherence using gpt-4. In **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications** *BEA*2023, 2023.

[5] Philippe Blache Israa Hamdine Lilia Azrou, Houda Oufaida. Using neural coherence models to assess discourse coherence. In **International Conference on Text, Speech, and Dialogue, 2024, Brno, Czech Republic**, 2024.

[6] 横野光. 誤用を含む文の形態素解析の分析とアノテーションの検討. 日本語学会 2024 年度春季大会 ワークショップ 2 “日本語学習者の作文コーパスの構築に向けて—メタ情報付与の方法—”, 2024.

[7] 文部科学省. 高等学校学習指導要領 (平成 30 年告示) 解説国語編. 東洋館出版社, 2019.

A　一貫性評価のプロンプトに使用した一貫性に関する知識

A.1　一貫性の説明

一貫性とは，「話題の焦点の遷移」といった要素からなる意味的関連性を指すとします．文章の筋が通っていてまとまりがあるかを評価してください．ここでは，一貫性と結束性を区別します．結束性は文法的関係や語彙的関係によって形成されます．結束作用の手段である接続表現の役割は既に存在する一貫性をさらに強めるという補助的役割を担っているものとします．この「既に存在する一貫性」の評価のために，文法事項とは切り離して一貫性を評価します．

A.2　判断理由から抽出した評価観点

- 書き手の言いたいことが伝わる
- 順序立ててまとめられている
- 一つのテーマに沿って書かれている
- 論理性に問題がない
- 主張にブレがない
- 唐突な転換や飛躍がない
- 言及のバランスに偏りがない
- 不要な語がない

A.3　指導要領における記述

『一般に，一つの段落には，中心となる一つの文と，その内容を支える (言い換えたり，例を挙げたりする) 文のみが含まれ，中心となる文が複数含まれることはない』
『各段落の中心文だけをつなげて読めば，文章全体の要旨が理解できるように構成されている』
『全ての根拠・論拠は適切か，根拠から導かれた結論は妥当か，飛躍や逸脱はないか，また論証に過不足はないか，当初の問いにきちんと対応した結論になっているかなど，様々な観点から論理の整合性と一貫性を検討，吟味する』
『全ての部分が，結論に向かう論証の中で明確な役割を持っているか』
『各文の抽象度による並びは適切か』