文学的表現に特化した 英日文学翻訳の評価データセットの構築に向けて

荒川花蓮　梶川怜恩　二宮崇　後藤功雄
愛媛大学
{arakawa@ai.cs.，reon@ai.cs.，ninomiya.takashi.mk@，goto.isao.fn@}ehime-u.ac.jp

概要

大規模言語モデル（LLM）の進歩により，機械翻訳の品質は大きく向上し，従来は困難とされてきた文学翻訳への応用も現実的になりつつある．しかし，既存の評価指標では，文体・語用・修辞といった文学的要素が翻訳後の文章においてどの程度維持されているかを個々に評価することが難しい．本研究では，英日文学翻訳の評価における新たな分析単位とラベリング基準を検討するため，事前に参照訳文に対して翻訳判断が現れる文節および単語レベルでのタグ付けを行った評価セットを試験的に構築する．提案する評価セットを用いて LLM による翻訳結果を分析したところ，文単位の単一スコア出力に留まる従来の評価指標では捉えにくい翻訳上の操作や誤りを体系的に可視化できることが確認された．

1　はじめに

大規模言語モデル（LLM）の発展により，機械翻訳の性能は年々向上しており，既存のベンチマークで主流であるニュースやマニュアル等のドメインにおいては，人手翻訳と遜色ない品質が報告される事例が増えている [1]．それに伴い，これまで専門性の高い翻訳タスクである文学翻訳に対しても，機械翻訳技術を適用する試みが活発化している [2]．実際，2023 および 2024 年の機械翻訳の共通タスクでは文学作品を対象としたタスクが設定された [**?**]．

翻訳品質の向上を議論する上で，適切な評価指標の存在は不可欠である．しかし既存の評価指標には，文学翻訳の質を評価するうえで以下の 2 つの課題がある．一つ目は，BLEU [3] や COMET [4] などの評価指標は，翻訳品質を単一のスコアとして出力するのみであり，翻訳文に含まれる文学的要素を個別に分離して評価することができないということがある．二つ目は，MQM [5] のような詳細な人手評価フレームワークも存在するが，その評価項目は主に一般的な翻訳タスクを対象に設計されており，文学翻訳の観点を反映できていないということである．そのため，既存の指標のみで文学翻訳の多面的な質を評価することは不適切である [**?**]．

本研究では，以上の背景を踏まえ，英日文学翻訳評価のための評価データセットを試験的に構築する．参照訳文に対して文節および単語レベルでのタグ付けを行うことで，文単位の評価よりも精緻な分析を可能とし，既存指標では捉えられない文学的な翻訳上の操作や傾向を体系的に把握することを目指す．また，提案する評価セットを用いて LLM の翻訳結果を分析し，本枠組みの有効性を検証する．

2　関連研究

従来，BLEU [3] や COMET [4]，MQM [5] といった評価指標が機械翻訳の品質評価に広く用いられてきた．BLEU は n-gram の一致度に基づくため，語彙や文構造の一致を捉えることには長けている．しかし，表現の多様性が高い文学翻訳においては，参照訳文との形式的な一致が少なく，文学翻訳特有の文体や比喩表現，話者の意図の変化など，微細な表現上の工夫をスコアに対して十分に反映することが難しい．COMET は意味的類似度に基づき，文の意味の忠実性を評価できるものの，文体的・語用的ニュアンスや修辞的効果の再現性を十分に捉えられない．また，人手による多次元評価フレームワークである MQM は，誤訳の種類や文体の適切さなどを評価できる一方で，LLM 翻訳と人間翻訳の微細な差異や文学的工夫の再現度を安定的かつ定量的に捉えることが難しい場合があると報告されている [6]．

こうした課題を背景に，近年は文学翻訳特有の文体や修辞，話者意図を考慮した新たな評価指標が提案され始めている．MAS-LitEval [7] では，「用語一貫性」「叙述視点一貫性」「文体一貫性」という 3 つの

エージェントで各観点を文単位で評価し，スコアリングする枠組みが提案された．また，LiTransProQA [8] では，LLM による質問応答形式で，文化的・修辞的要素や著者の声がどの程度を保持されているかを文単位で評価する．さらに英-韓文学翻訳評価用の詳細アノテーション手法により，微細な文体や語用の変化まで捉えるデータセットの構築も取り組まれている [9].

3　提案する評価データセットの特徴

本研究が提案する評価データセットの特徴は，参照訳文において翻訳者の意図的な工夫（操作）が反映された特定の言語的スパンに対し，あらかじめタグ付けを行っている点にある．これにより，評価時に LLM がその翻訳操作をどの程度再現できているかを詳細に分析することが可能となる．既存の評価フレームワークである MQM が，評価者に誤訳箇所の特定と分類を委ねる「エラー分析」を主眼に置くのに対し，本手法では，事前に原文・参照文比較に基づき，翻訳者の工夫が顕著な「翻訳操作単位」をあらかじめ評価対象として指定する．このアプローチにより，特定の翻訳機能（例：比喩の保持，主語の明示化）に焦点を当てた分析が可能になるだけでなく，評価箇所が固定されるため，評価者間の判定の揺れを抑制し，評価の客観性を高めることができると考えられる．また，MAS-LitEval や LiTransProQA が参照文を用いない自動評価や，段落・章レベルでの大局的なスコアリングを目指すのに対し，本研究は参照訳文を基盤とした微細なスパンレベルの分析に立脚している．これにより，BLEU や COMET のような全体的な類似度スコアでは埋没しがちな，「比喩が適切に変換されているか」「文脈に応じた明示化が行われているか」といった，文学翻訳の質を決定付ける具体的な機能ごとの精密な評価が実現される．

4　評価データセットの構築方法

4.1　データセットの前処理

構築元のデータとして，NICT が作成した日英対訳文対応付けデータ [10] を利用する．これは，以下のサイトで公開されている文学作品などを文単位で対応付けしたものである．

- Gutenberg[11]：英語の原文，翻訳文
- 青空文庫 [12]：日本語の原文，翻訳文
- 杉田玄白プロジェクト [13]：有志による英語原文の日本語訳

またデータ内には，作品本文だけでなくタイトル，作者，脚注なども含まれており，文対の数は 12,031 である．本研究では，評価対象モデルの文学的な翻訳工夫，表現能力を多角的に評価することのできる評価セットを目標とし，以下の処理を施す．

- 小説以外の作品を除外し，英語を原文とする文学作品のみを抽出する
- 翻訳の質を担保するために，過去に翻訳本の商業出版実績がある翻訳者による作品のみを採用する
- 物語の連続性や文脈を利用した翻訳判断（指示語の解釈など）を評価するために，各作品の先頭から順に，訓練，検証，テストを 7 : 1.5 : 1.5 の比率で分割する．本研究では，この検証セットおよび一部のテストセットを分析対象とする．

4.2　検証セット分析による翻訳判断の体系化

翻訳手続きの体系化 [14] による翻訳操作の区別をもとに，検証セットの原文と参照訳文を分析し，プロの翻訳者がどのような目的で，どのような部分に対して工夫を凝らしたかという「翻訳判断」を，以下の 4 つの主要カテゴリに体系化した．

翻訳判断の主要カテゴリ

情報量操作：文脈情報を利用し，指示語の明示化や単語などの補完，原文にある要素をあえて訳出しないなど，読者の理解を助けるための言語的配慮

語用・発話態度：話し手の意図，口調や人称や発話内に含まれるキャラクターの役割語で，登場人物間の人間関係や感情の機微を表現し，作品の世界観やリアリティを構築

表現運用：同じ語句の言い換え，説明的修飾部を複合名詞に言い換えるといった，可読性や文章の引き締まりを追求し，文学的文脈やジャンルに即した最適な語彙を選択・適用

文学的効果・修辞：芸術性や臨場感，比喩や直喩といった表現技法で，文学翻訳特有の読者の感性に訴えかけるための技巧的な工夫

表 1　翻訳者の意図が反映された最小言語単位（単語・文節）へのタグ付け事例

カテゴリ	原文（該当箇所）	参照訳文（付与単位）	翻訳操作の概要
情報量操作	I have done so	加害者も**捕えて**	文脈に基づく不足情報の補完（代動詞の明示化）
語用・発話態度	or why should he...	**でなければ何んで**	文脈の論理矛盾を突く反語の再現
表現運用	real murderer	**真犯人**	説明的表現の定型的な熟語への凝縮
文学的効果・修辞	The horse!	馬がッ！	促音による驚愕の音声的な演出

表 2　評価事例：入力文量による翻訳品質の比較（プロ訳・1 文入力の LLM 訳・5 文入力の LLM 訳の対比）

カテゴリ	翻訳の種類	翻訳文	評価
情報量操作	プロ訳	加害者も**捕えておきました**	-
	LLM(1 文)	**そうしました**	×
	LLM(5 文)	もう**捕まえましたよ**	○
	プロ訳	**ホームズ**は	-
	LLM(1 文)	**シャーロック・ホームズ**は	×
	LLM(5 文)	**シャーロック・ホームズ**は	×
語用・発話意図	プロ訳	**でなければ何んで**薬で厩番を眠らせ**たりしましょう.**	-
	LLM(1 文)	**それとも，なぜ**彼が自分の厩舎係の少年に薬を盛ら**なければならないんだ？**	△
	LLM(5 文)	**でなければ，なぜ**自分の厩務見習いに薬を盛る**必要がある？**	○
	プロ訳	どこ**です**？	-
	LLM(1 文)	どこ？	×
	LLM(5 文)	どこ**だ**？	△
	プロ訳	えッ！　**捕まったって**？	-
	LLM(1 文)	あいつを**捕まえたぞ！**	×
	LLM(5 文)	**捕まえたのね！**	×
表現運用	プロ訳	**真犯人**はあなたのすぐ後に立っていますよ	-
	LLM(1 文)	**本当の犯人**は，今まさにあなたのすぐ後ろに立っている.	△
	LLM(5 文)	**本当の殺人者**は，あなたのすぐ背後に立っています.	×
	プロ訳	大佐，あなたの**競馬に関する広い経験**をもってすれば，	-
	LLM(1 文)	**競馬のことにはご経験が豊富な**ロス大佐ならご存じのはずですが，	△
	LLM(5 文)	**芝の世界に関して豊富なご経験をお持ちの**ロス大佐ならご存じでしょうが	×
文学的効果・修辞	プロ訳	**まるで**傍で見ていた**よう**です！	-
	LLM(1 文)	君もそこにいたかもしれない！	×
	LLM(5 文)	君もそこにいたかもしれない！	×

4.3　テストセットの参照訳文の翻訳判断部分に対してタグ付け

検証セットにおける分析に基づき，体系化したカテゴリを用いて，日本語を母国語とする大学生一人が，評価セットの参照文 100 文に対し，人手でタグ付けを行った．各タグは，翻訳者の特定の意図が反映された最小の言語単位（文節や単語レベル）に対して付与されている．実際にタグ付けされた例を表 1に示す．また，タグ付けの結果 100 件の評価セットを得た．

5　評価実験

提案する評価セットの有効性を検証するため，LLM を用いて評価セットを翻訳した．

5.1　実験設定

モデル　翻訳モデルは gpt-5.2-2025-12-11 を採用し，推論時の温度パラメータは 0.0 に設定した．本実験では，以下のようなプロンプトを与えた．

System Prompt
You are a professional literary translator (English → Japanese), specializing in narrative and dialogue.
Preserve tone, style, and nuance.
Output ONLY a valid JSON object in the following format:
{"translation": "..."}
Do not include explanations or notes.
User Input
Translate the following sentence: [原文]

また，出力品質の一貫性向上のため，出力形式を JSON 形式とした [15].

5.2 評価事例

実験により得られた表2を用いて，各カテゴリで以下のような考察を行った．

情報量操作 明示化において，入力文量の違いによる「文脈依存性」が顕著に現れた．

- **明示化の成功と失敗：** 原文の代動詞に対し，1文入力では「そうしました」と指示対象が不明瞭な直訳にとどまった．一方で，5文入力では，前後の文脈より「犯人を捕まえた」という事実を特定し，動詞を明示化することに成功した．
- **暗黙化（省略）の困難さ：** LLM訳では，原文に存在する固有名詞を忠実に訳出する傾向がみられた．だが，日本語では主語として固有名詞を利用する際，フルネームを利用することは多くなく，プロ訳では「シャーロック」が省略された．

語用・発話態度 文脈の増加に伴い，キャラクター性が強調される一方で，属性の誤認や発話意図の変質が生じる傾向が確認された．

- **キャラクター口調の深化：** 1文入力のLLM出力では「どこ？」といったフランクな口調であり，5文入力では，「どこだ？」と断定的かつ男性的な口調であった．しかし，プロ訳では「どこです？」と丁寧な口調だった．
- **口調による性別の変化と発話意図：** 5文入力では「〜のね！」という女性的な口調となったが，発話者は男性であった．プロ訳では，「(もう犯人を）捕まえたって？」という驚き・疑問が含まれていると解釈できるが，LLM出力では，「〜ぞ！」「〜のね！」により驚きのみ強調され，発話意図が変質した．
- **反語の変化：** 1文入力では「それとも」という文頭と，「なぜ」という翻訳により，疑問としての面が強い．一方5文入力では，文頭が「でなければ」であるため，「(前の内容）でないならなぜBをする必要があるのか？（いや，ない)」と解釈でき，反語として機能した．

表現運用 表現の言い換えはおおむね可能であったが，説明的表現を端的に表現するような，可読性を高める操作（短縮）は困難であった．

- **言い換え表現：** プロ訳は，「経験」を「競馬に関する」「広い」という語が修飾する場合に，1文入力では，プロ訳の「広い」と同じ意味で「豊富な」が利用されたが，全体を通して日本語として不自然な表現となった．5文入力では，プロ訳の「競馬に関する」の言い換えが「芝生の」にという誤りが起こった．
- **複合名詞への変換：** 原文の “The real murderer” に対し，プロ訳はミステリの定型表現である「真犯人」という語を選択した．これは，説明的な要素を削ぎ落として読者に強い印象を与える表現運用（短縮）だといえる．対してLLMの訳文では，1文入力が「本当の犯人」，5文入力が「本当の殺人者」と表現された．特に5文入力は，原文の “murderer” が忠実な訳出だ．しかし，日本の探偵小説において，解決編でのこの呼称は不自然であり，文学的慣習よりも字義通りの正確さを優先するLLMの問題点が示された．

文学的効果・修辞 場面の臨場感を演出する直喩や暗喩といった表現技法の保持はできたが，修辞の意図的な変換は困難であった．

- **比喩の解釈：** 参照訳文では，文化圏の違いを考慮し，暗喩を「まるで傍で見ていたよう」と直喩的にわかりやすく変換した．対して，LLMは「君はそこにいたかもしれない」と暗喩を事実的に翻訳し，文学的な技巧が損失した．

6 おわりに

本研究では，英日文学翻訳の品質評価における新たなアプローチとして，事前に参照文に対して翻訳判断がみられる文節・単語レベルでのアノテーションを施した評価セットを試験的に構築した．評価実験の結果，既存指標では捉えきれなかった，LLMによる翻訳の微細な傾向が明らかとなった．具体的には，「情報量操作」「語用・発話意図」「表現運用」「文学的効果・修辞」それぞれについて，LLMによる翻訳の特徴がみられた．本手法は，「どの機能的側面で翻訳が不十分か」を特定箇所と紐づけて提示できる点に優位性があり，多角的な文学翻訳の品質評価を可能にする．今後は，テストセット全件まで評価セットの拡充を進める．同時に，本評価セットを正解データとして活用し，LLM-as-a-judgeによる自動評価が文学的な差異をどの程度正確に識別できるかについても検証したい．最終的には，英日文学翻訳に特化したベンチマークの確立を目指す．

謝辞

本研究は国立研究開発法人情報通信研究機構の委託研究（課題番号：225）および JSPS 科研費 JP24K15071 の助成を受けたものです.

参考文献

[1] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In **Proc. of WMT**, pp. 1–46, 2024.

[2] Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. Exploring Document-Level Literary Machine Translation With Parallel Paragraphs From World Literature. In **Proc. of EMNLP**, pp. 9882–9902, 2022.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In **Proc. of ACL**, pp. 311–318, 2002.

[4] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In **Proc. of EMNLP**, pp. 2685–2702, 2020.

[5] Andreas Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. **Tradumàtica: International Journal of Translation and Translation Technology**, Vol. 12, pp. 455–467, 2014.

[6] Ran Zhang, Wei Zhao, and Steffen Eger. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In **Proc. of NAACL**, pp. 10961–10988. Association for Computational Linguistics, 2025.

[7] Junghwan Kim, Kieun Park, Sohee Park, Hyunggug Kim, and Bongwon Suh. MAS-LitEval : Multi-Agent System for Literary Translation Quality Assessment. **arXiv:2506.14199**, 2025.

[8] Ran Zhang, Wei Zhao, Lieve Macken, and Steffen Eger. LiTransProQA: An LLM-based Literary Translation Evaluation Metric With Professional Question Answering. In **Proc. of EMNLP**, pp. 29087–29109, 2025.

[9] Sheikh Shafayat, Dongkeun Yoon, Jiwoo Choi, Woori Jang, and Seohyon Jung. Evaluating English-Korean Literary Machine Translations: A Dataset Featuring the RULER and VERSE Annotation Methods. **JOHD**, Vol. 11, , 2025.

[10] 内山将夫, 高橋真弓. 日英対訳文対応付けデータ. https://www2.nict.go.jp/astrec-att/member/mutiyama/align/index.html, 2003. 公開データセット.

[11] Project Gutenberg. Project gutenberg. https://www.gutenberg.org/, 2025. 最終アクセス日: 2025-01-08.

[12] 青空文庫. 青空文庫. https://www.aozora.gr.jp/, 2025. 最終アクセス日: 2025-01-08.

[13] プロジェクト杉田玄白. プロジェクト杉田玄白. https://genpaku.org/, 2025. 最終アクセス日: 2025-01-08.

[14] Jordi Tronch. Translation procedures. https://www.uv.es/tronch/Tra/TranslationProcedures.html, 2025. 最終アクセス日: 2025-01-09.

[15] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does Prompt Formatting Have Any Impact on LLM Performance? **arXiv:2411.10541**, 2024.