視覚言語モデルは漫画のオノマトペを翻訳できるか？

濱本惇之介[1]　梶川怜恩[1]　二宮崇[1]　後藤功雄[1]　石渡祥之佑[2]　能地宏[2]
[1]愛媛大学　[2]Mantra 株式会社
{hamamoto@ai.cs., reon@ai.cs., ninomiya.takashi.mk@}ehime-u.ac.jp
goto.isao.fn@ehime-u.ac.jp,{ishiwatari,noji}@mantra.co.jp

概要

漫画のオノマトペは，音や感情，動きなど漫画の場面を視覚的に補完する重要な要素だが，翻訳先の多様性および視覚的文脈への依存ゆえに，従来の機械翻訳では十分に検討されてこなかった．本研究では，視覚言語モデル（VLM）を用いた漫画のオノマトペ翻訳の有効性を検証する．実験の結果，VLM は視覚情報の利用により，翻訳精度を向上させることが確認された．一方で，人手翻訳に比べ語彙の多様性に欠けるという課題も明らかになった．加えて，既存の評価指標の限界を示し，本タスク特有の評価尺度の必要性について述べる．

1　はじめに

日本の漫画は世界中で人気を博しており，多くの作品が翻訳され海外読者にも親しまれている [1]．漫画の魅力を構成する要素の一つとしてオノマトペ [2] が挙げられる．オノマトペは擬音語および擬態語から構成される語で，音や感情，状態などを臨場感豊かに表現できる．そのため，図 1 のように適切に翻訳することができれば海外の読者にも原作と同じ体験を届けることが可能である．

しかし，オノマトペの翻訳にはいくつかの課題が存在する．第一に，翻訳表現の多様性がある．同一の状況・同一の意味でも，翻訳者により表現が大きく異なる場合 [3] がある．第二に，視覚的文脈依存性が高い点である．オノマトペは音や雰囲気を主に表すため意味の幅が広く，同じ表記であっても文脈によって訳語が大きく変化する．例えば，日本語の「ぱたぱた」は状況によって訳語が異なる．旗が揺れる描写では “flap”，廊下を走る時の足音では “pitter-patter” と訳されるなど [2] 文脈に応じた訳し分けが必要となる．そのため，適切な翻訳を行うには深い状況理解が不可欠となる．

さらに，漫画特有の問題として，状況を詳述する

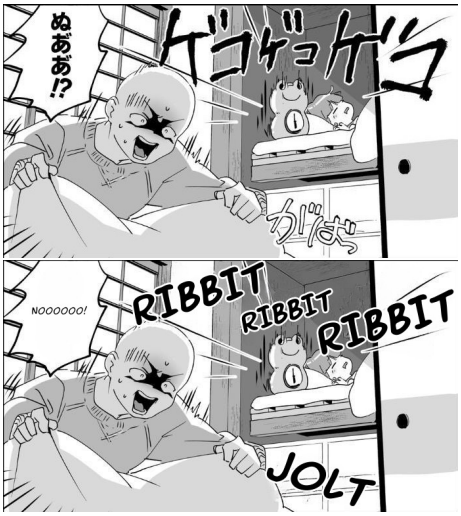

図 1: 漫画のオノマトペ「ゲコ」と翻訳結果 “RIBBIT”（カエルの鳴き声）を示した作例　© なめこなこ

情報が翻訳対象のテキストから読み取れない場合が多いことが挙げられる [4–8]．漫画では状況がテキストではなく画像で表される．そのため，文字情報のみで文脈を捉えることは困難なケースが少なくない．これに対し，視覚言語モデル（VLM）で画像情報を補完し，文脈理解を改善する研究 [6, 8] があるが，多くは吹き出し内のセリフが対象でオノマトペについては十分に検討されていない．

そこで本稿では，VLM を用いて漫画内のオノマトペを翻訳し，その有効性と課題を分析することを目的とする．分析の結果，既存手法によるオノマトペ翻訳の精度改善が確認された．一方で，生成される語彙の多様性の欠如や，既存の評価指標における限界が浮き彫りとなった．本研究の成果は，VLM を用いたオノマトペ翻訳における今後の研究指針となることが期待される．

2　関連研究

漫画のオノマトペの特徴　日本の漫画におけるオノマトペの特徴として，非常に創造性に富み，絶

表 1: Few-shot 条件の違いによる翻訳性能の比較

	zero-shot			3-shot				5-shot
	no-image	image	image+genre	random	neologism	same-genre	cross-genre	random
chrF	11.14	11.56	12.98	13.57	10.95	13.30	12.47	14.00
COMET	0.5857	0.5876	0.5924	0.5901	0.5854	0.5913	0.5854	0.5901

えず新たな表現が生み出されている点が挙げられる [9]．英語のコミックにおいても同様の傾向がみられ，造語や独自のスペリングの使用が確認されている [10]．漫画のオノマトペ翻訳の傾向を分析した研究 [11, 12] では，こうした特徴が翻訳においても現れ，オノマトペの翻訳手法として「造語」が多用されていることが指摘されている．例えば，風音の「ヒューッ」を文字の重ねで表す “FUUUUIIIIISH” や，扉の音「キッ」に独自のスペリングを用いた “KWIII” などが挙げられる．

漫画の機械翻訳　漫画の吹き出しに対する機械翻訳の精度向上を目的として，テキスト以外の補助情報を活用するアプローチが提案されている．例えば，作品のジャンル情報などの属性情報を入力として付加する手法 [5] や，VLM を用いて漫画画像をテキストと同時に入力する手法 [6] などが報告されており，いずれも文脈理解における補助情報の有効性を示している．一方，オノマトペの機械翻訳に関する既存研究は少なく，Google 翻訳を用いた翻訳傾向の検証 [13] があるが，テキスト入力のみを対象とした分析に留まっている．そのため，吹き出し翻訳で有効とされる補助情報が，オノマトペ翻訳でも同様に有効であるかは明らかになっていない．本研究では，視覚情報やジャンル情報がオノマトペ翻訳に与える影響と有効性を検証する．

3　実験

データセット　本研究では，先行研究において構築された日英の漫画画像からなる Manga Corpus [4] を基に，評価および検証に用いるデータセットを構築した．Manga Corpus のうち，オノマトペが英語に翻訳されている作品を対象とし，オノマトペを含むコマ画像と，それに対応する日英の対訳対をそれぞれ人手で抽出した．抽出したデータのうち，評価データとして，10 作品から各 100 ペア，計 1,000 件を構築した．さらに few-shot に用いる検証データとして，評価データと同一の 10 作品から 20 ペアずつ抽出した作品内検証データ，および評価用データには含まれない別の 10 作品から 20 ペアずつ抽出した作品外検証データを構築し，計 400 件の検証データを作成した．

実験設定　本研究では，VLM のモデルとして Qwen3-VL-30B-A3B-Instruct[1] [14] を使用した．モデルの推論時のハイパーパラメータは Temperature を 0，最大生成トークン数は 128 に設定した．モデルへの入力は，日本語オノマトペのテキスト，およびそれに対応する漫画のコマ画像である．本実験では，これらを翻訳指示を含むプロンプト[2]に埋め込み，英語のオノマトペを生成させた．

評価方法　評価時には，モデルの出力に対して翻訳箇所の抽出および不要な記号の除去等の正規化を行った．その後，参照訳とともに英小文字化し，評価を行った．評価指標としては，表層的な類似度を測る chrF [15] と，意味的な類似度を測る COMET [16][3]の 2 つを採用した．

4　実験結果

4.1　定量的分析

表 1 に示す実験結果に基づき，入力条件の違いが翻訳性能に与える影響について分析する．

視覚情報およびジャンル情報の有無による影響　zero-shot 設定において，画像の有無およびジャンル情報が翻訳に与える影響を検証した．比較条件として，翻訳対象のオノマトペのみを入力とする条件（no-image），テキストに加え対応するコマ画像を入力とする条件（image），さらに先行研究 [5] に基づき，コマ画像に加えて対象作品のジャンル情報をプロンプトに付加する条件（image+genre）を設定した．なお，ジャンル情報の取得には各作品に対応する Wikipedia の Infobox および漫画専門のオンライン百科事典であるマンガペディア[4]の記載を参照した．

実験結果として，no-image と image を比較すると，chrF および COMET が一貫して上昇しており，翻訳

1）https://huggingface.co/Qwen3-VL-30B-A3B-Instruct
2）モデルに与えるプロンプトは付録 A に示す．
3）https://huggingface.co/Unbabel/wmt22-comet-da
4）https://mangapedia.com/

対象テキストにコマ画像を加えることが翻訳性能向上に有効であることが示された．さらに，image と image+genre を比較すると，両指標ともにさらなる向上が確認された．このことから，オノマトペ翻訳においても視覚的文脈に加え，ジャンル情報の明示が精度改善に寄与すると示された．

提示事例数の違いによる影響　few-shot 設定において，モデルに提示する入力事例の数が翻訳性能に与える影響を検証した．本実験では，作品内検証データから翻訳対象と同一の作品に含まれる翻訳対および対応するコマ画像を無作為に抽出し，これらを入力例とする条件（random）を設定した．この条件下で，3-shot および 5-shot の推論を行った．

zero-shot (image), 3-shot (random), 5-shot (random) の結果を比較すると，chrF は事例数の増加に伴い一貫して向上した．COMET については 3-shot と 5-shot で同程度の値となったものの，zero-shot を上回る結果となった．このことから，オノマトペ翻訳においても，適切な入力事例の提示が翻訳性能の向上に寄与することが示された．

入力事例に含まれる造語の影響　few-shot の入力例として，オノマトペ特有の「造語」が含まれる事例の影響を分析した．先行研究 [11] では造語の重要性が指摘されていることから，本実験では作品内検証データより参照訳に造語を含む事例を抽出した．実験では，3-shot の例として造語を含む事例のみを与えた場合（neologism）と無作為に事例を与えた場合（random）との比較を行った．

実験の結果，random と neologism を比較すると，neologism 条件において chrF および COMET は一貫して低下しており，造語が含まれる事例を与えた場合に翻訳精度が悪化する傾向が見られた．漫画のオノマトペの特徴である造語を入力例として与えることで翻訳精度の向上が期待されたが，実際には精度は大きく低下した．この結果から，作品全体で見ると造語の使用頻度は低いため，特殊な事例である造語を入力例として用いたことが，一般的なオノマトペの翻訳においてノイズとなり，翻訳精度の低下を招いたことが示唆される．

異なる作品からの事例提示におけるジャンルの影響　翻訳対象とは異なる作品から事例を提示する場合のジャンルの影響を分析した．作品外検証データを用い，翻訳対象と同じジャンルの作品から無作為に事例を抽出した条件（same-genre）と異なるジャンルの作品から無作為に抽出した条件（cross-genre）を設定し，それぞれ 3-shot で比較を行った．

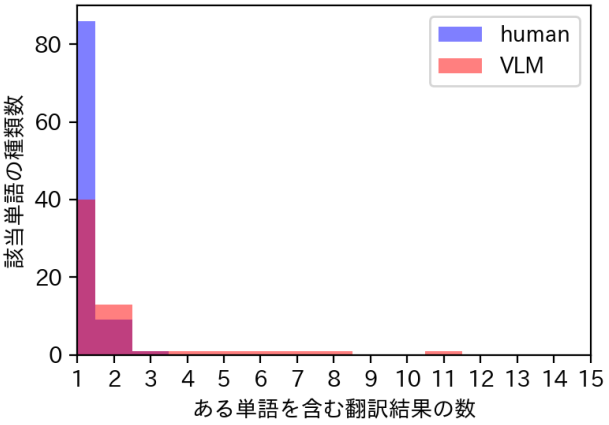

図 2: 人間の翻訳結果と VLM の翻訳結果における単語の使用頻度分布

実験の結果，他作品の事例を与えた場合（same-genre, cross-genre）は，同一作品の事例を与えた場合（random）と比較して，chrF が全体的に低下した．しかし内訳を見ると，同ジャンル（same-genre）では chrF の低下はわずかであり，COMET に関しては random を上回る精度を示した．対照的に，異なるジャンル（cross-genre）を用いた場合は両指標ともに大きく悪化した．これは，ジャンルごとにオノマトペ表現の傾向や特徴が共通していることを示唆している．すなわち，同一作品の事例が利用できない場合でも，同ジャンルの事例であれば，翻訳品質を維持，あるいは意味的な妥当性を向上させられることが明らかになった．

4.2　統計的分析

VLM によるオノマトペ翻訳における語彙選択の傾向を分析するため，画像入力を行った zero-shot の翻訳結果と，評価データの参照訳に含まれる単語の出現頻度を比較した．作品ごとに人手翻訳と VLM の出力語彙を抽出し，100 件の翻訳結果を対象として，各単語の出現頻度を算出した．図 2 はある作品（アクション漫画）における単語の出現頻度を横軸，該当する単語数を縦軸として作成したヒストグラムである．

分析の結果，人手翻訳では，出現する語の約 8 割が 1 回のみの出現であり，語彙の多様性が高い分布となった．一方，VLM の出力では，1 回のみ出現する語の割合は約 4 割まで低下し，多くの語が 2 回以上繰り返し使用されていた．さらに，図 2 の例では，100 件中 11 件に同じ単語が使われるなど，語彙

の使い回しが顕著に見られた．なお，この傾向は他作品においても同様に確認された．

次に，VLM の出力において特に出現頻度の高い語に着目し，それらがどのような場面で出力されているのか，人手翻訳との比較を通じて分析を行った．分析対象として，図 2 の作品において VLM が最も多用した，重い衝撃音を表す “thud” を取り上げる．まず，ものを殴る描写で用いられた打撃音の翻訳事例を確認すると，人手翻訳では「ドカ」に対して “wham”,「ドス」に対して “thud”,「ドス ドス」に対して “tump tump” と，衝撃の音や回数に応じた多様な語彙選択が行われていた．一方で，VLM はこれらの異なるオノマトペをすべて一律に “thud” と翻訳しており，表現の多様性に乏しいことが明らかとなった．

さらなる問題点として，お辞儀の場面における「ペコ ペコ」や，鐘の音を表す「カーン」に対しても “thud” が出力される事例が見られた．これらは衝撃音とは無関係な描写であり，VLM が画像内の文脈を無視し，汎用的な高頻度語を出力している様子が確認された．この傾向は，他の作品や頻出語においても共通して確認された．これらの結果から，VLM は人手翻訳に見られるような，場面に応じた語彙の多様な使い分けが十分に行われていないことが明らかになった．

4.3 定性的分析

画像の有無による翻訳結果の比較 オノマトペ翻訳における画像情報の影響を分析するため，zero-shot で，オノマトペのテキストのみを入力した場合と，それに加えて対応するコマ画像を入力した場合の翻訳結果を比較した．評価データセット内で同一の日本語オノマトペに対する翻訳結果の差異を人手で分析した結果，画像情報を併用することにより，翻訳結果が改善される事例が複数確認された．

例えば，少女が泣く場面の「ボロ ボロ」に対して，画像を入力しない場合には，「がたがた」といった物理的な音を表す “rattle rattle” と翻訳された．一方，対応するコマ画像を入力として与えた場合には，泣いている様子を表す “sobbing weeping” と翻訳され，文脈に即した解釈が行われていることが確認された．テキスト情報のみでは解釈が困難となる場合において，視覚情報の併用が妥当な訳語生成に寄与したことが確認され，漫画のオノマトペ翻訳における視覚情報の有効性が示された．

評価指標の限界 オノマトペ翻訳に対する評価指標の評価傾向を分析するため，画像入力を行った zero-shot の翻訳結果のうち，chrF および COMET のスコアが低い事例下位 10%を抽出し，参照訳との比較分析を行った．

まず chrF に関しては，翻訳の妥当性にかかわらず，表層的な不一致によりスコアが著しく低くなる事例が散見された．具体的には，強調表現の「ドーン」に対して参照訳が “DOOM”，VLM 出力が “BANG” となった事例や，うなされる場面の「んー」に対して参照訳が “unn”，VLM 出力が “hmm” となった事例が挙げられる．いずれも意味的には許容可能であり，表現としても自然であるにもかかわらず，文字列が一致しないために chrF では評価対象とならず，スコアが 極めて低い値となった．

次に COMET に関しても，翻訳結果が文脈に適しているが，不当に低く評価される事例が確認された．具体的には，電話の呼び出し音「トゥルルルルルルッ」に対して参照訳が “riiiiing”， VLM 出力が “ring ring ring ring ring ring!” となった事例や，ドアを強く叩く「ダン ダン ダン」に対して参照訳が “knock knock knock”， VLM 出力が “dun dun dun” となった事例が挙げられる．これらは着信音やノック音の表現として自然であり，文脈上も妥当だが，COMET スコアは低い値にとどまった．この要因として，COMET が，オノマトペ特有の表現を十分に学習していない可能性が高い．長音化や音の響きを模倣する表現は一般的なテキストコーパスには乏しいため，埋め込み空間上で参照訳との意味的等価性が正しく捉えられず，スコアの過小評価につながったと考えられる．これらの結果から，オノマトペ翻訳の性能評価においては，既存の評価指標をそのまま適用することの限界が示唆された．

5 おわりに

本研究では，VLM による漫画オノマトペ翻訳の有用性と課題を分析した．実験の結果，画像やジャンル情報の利用が翻訳精度の向上に有効であることが示された．一方で，VLM は人手翻訳に比べ語彙の多様性に乏しく，文脈を無視した汎用的な訳語を多用する傾向が明らかとなった．また，既存の評価指標はオノマトペ特有の表現を適切に捉えきれず，人間の感覚と乖離する事例も確認された．今後は，オノマトペ翻訳の品質をより正確に測定できる評価指標の構築に取り組む．

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] 日本経済新聞. 漫画を AI で多言語翻訳、迅速輸出で海賊版を防止 文化庁が人材育成, 2026. https://www.nikkei.com/article/DGXZQOUD200JQ0Q5A221C2000000/.

[2] 吉村耕治. 日英語の比較の観点から見たオノマトペ : 感性の表現の魅力. 表現研究 / 表現学会 編, Vol. 102, pp. 7–18, 2015.

[3] Azari Raziyeh and Sharififar Masoud. Translating onomatopoeia: An attempt toward translation strategies. **Review of Applied Linguistics Research**, Vol. 3, No. 3, pp. 72–92, 2017.

[4] Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. Towards Fully Automated Manga Translation. In **Proceedings of the AAAI Conference on Artificial Intelligence**, 2021.

[5] Hiroto Kaino, Soichiro Sugihara, Tomoyuki Kajiwara, Takashi Ninomiya, Joshua B Tanner, and Shonosuke Ishiwatari. Utilizing longer context than speech bubbles in automated manga translation. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation**, 2024.

[6] Philip Lippmann, Konrad Skublicki, Joshua Tanner, Shonosuke Ishiwatari, and Jie Yang. Context-informed machine translation of manga using multimodal large language models. In **Proceedings of the 31st International Conference on Computational Linguistics**, 2025.

[7] 眞鍋光汰, 梶原智之, 二宮崇, 後藤功雄, 石渡祥之佑, 能地宏. マルチモーダル漫画翻訳のための画像エンコーダのドメイン適応. 情報処理学会 自然言語処理研究会, 2025.

[8] 横山泰知, 戒能大翔, 梶川怜恩, 二宮崇, 後藤功雄, 石渡祥之佑, 能地宏. 場面説明を活用したマルチモーダル漫画翻訳. 言語処理学会第 32 回年次大会発表論文集, 2026. To Appear.

[9] 山口仲美. 犬は「びよ」と鳴いていた：日本語は擬音語・擬態語が面白い. 光文社, 2002.

[10] エリノア・H・ジョーデン. 擬声語・擬態語と英語. 日英語比較講座 4 発想と表現. 大修館, 1982.

[11] Inose Hiroko. Translating japanese onomatopoeias and mimetic words in manga. **Japan Association for Interpreting andTranslation Studies**, No. 10, pp. 161–176, 2010.

[12] Martin Teshome. Onomatopoeia in a japanese-to-english translation of all out!!: a case study. 日本文理大学紀要 (Bulletin of Nippon Bunri University), Vol. 52, No. 1, pp. 17–26, 2024.

[13] Richard Reenald and Elisa Carolina Marion. Analysis of onomatopeia translation result on "komi can’t communicate" comic using machine translation. In **Proceedings of the 3rd Asia Pacific International Conference on Industrial Engineering and Operations Managemen**, 2022.

[14] Qwen Team. Qwen3-VL Technical Report. Technical report, arXiv, 2025.

[15] Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In **Proceedings of the Tenth Workshop on Statistical Machine Translation**, 2015.

[16] Rei Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In **Proceedings of the Seventh Conference on Machine Translation (WMT)**, 2022.

```
Please look at the image and translate the Japanese sound effect '[Input Onomatopoeia]' into English.
Consider the visual context to choose an appropriate sound effect that captures the impact, motion, and
tone shown in the image.
Just give the best English equivalent sound effect or onomatopoeia, without explanation.
```

図 3: オノマトペ翻訳のプロンプト例

```
This manga's genres are: [Genre Tags].
Please look at the image and translate the Japanese sound effect '[Input Onomatopoeia]' into English.
Consider the visual context, and take the manga's genres into account, to choose an appropriate sound
effect that captures the impact, motion, and tone shown in the image.
Just give the best English equivalent sound effect or onomatopoeia, without explanation.
```

図 4: ジャンル情報を付与したプロンプト例 (赤字は図 3 との差分)

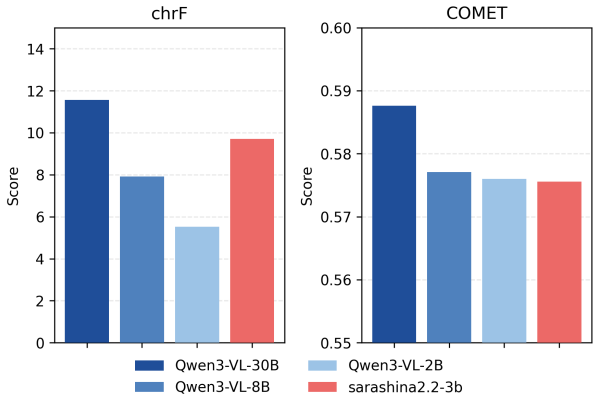

図 5: モデルおよびパラメータの違いによる翻訳性能の比較

付録

A　入力プロンプト

本実験で用いた基本プロンプトを図 3 に示す．また，4.1 節の分析において，ジャンル情報を付加した実験で用いたプロンプトを図 4 に示す．プロンプト内の赤字部分はジャンル情報を付加する際に追加した箇所である．

B　モデル・パラメータの比較

モデルおよびパラメータの違いが翻訳性能に与える影響を検証した．比較対象として，Qwen3-VL-30B に加えて，Qwen3-VL-8B[5]，Qwen3-VL-2B[6]，および sarashina2.2-3B[7] を用いた．実験は zero-shot 設定で行い，翻訳対象のオノマトペと対応するコマ画像をモデルに入力した．

実験結果を図 5 に示す．実験結果より，Qwen3-VL の 2B,8B,30B を比較するとパラメータ数の増加に伴い chrF および COMET が一貫して上昇した．このことから，パラメータ数の増加が翻訳の性能向上に寄与することが示された．

また，sarashina2.2-3b を Qwen3-VL の各モデルと比較すると，chrF では Qwen3-VL-8B と 30B の中間に位置したものの，COMET では最低値にとどまった．出力結果を確認したところ，sarashina2.2-3b には Qwen3-VL と比較してオノマトペをローマ字転写する事例が多く確認された．ローマ字表記自体は翻訳戦略として妥当であるが，意味的な等価性を評価する COMET においては，ローマ字転写による翻訳は意味内容を十分に反映した表現としてみなされにくく，参照訳との意味的整合性が低いと判断される．その結果，文字列一致を重視する chrF と比較して著しく低い評価となったと推察する．

5) https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
6) https://huggingface.co/Qwen/Qwen3-VL-2B-Instruct
7) https://huggingface.co/sbintuitions/sarashina2.2-3b