Buletin Edukasi Indonesia (BEI)
DOI: 10.56741/IISTR.bei.002724

E-ISSN 2962-1674
P-ISSN 2962-5742



Human and ChatGPT Translations of José Rizal's Poetry

A Qualitative Comparison

[1]Christopher A. Britanico, [1]Nelia A. Britanico, [1]Maricon V. Quinto, [1]Marissa Del Rosario
Corresponding Author: * cbritanico@usant.edu.ph
[1] University of Saint Anthony, Iriga City, Philippines

Article submitted: 28 September 2026
Revised: 13 October 2026
Accepted: 18 October 2026

Abstract

This qualitative corpus-based study compares selected passages from José Rizal's Spanish poems "Mi Último Adiós" and "A la juventud filipina" with human-produced English and Tagalog translations and ChatGPT outputs documented in an earlier manuscript. Guided by translation-shift analysis, reflexive thematic interpretation, and critical discourse analysis, it examines semantic transfer, metaphor, cultural references, ideological framing, and poetic form. The source manuscript reports several illustrative instances in which ChatGPT retains propositional meaning but simplifies metaphor or changes the register of patriotic expressions. The evidence is interpreted as a small, purposively selected set of examples rather than as proof of model-wide failure. The revised analysis distinguishes observable textual differences from claims requiring additional documentation, including the identities and editions of human translations, the exact ChatGPT model, generation dates, and independently checked source-text alignments. Findings support cautious, context-dependent use of AI as a drafting aid alongside historically informed human review. The article offers a transparent coding framework and a reproducibility checklist for future multilingual literary-translation research.

Keywords: *ChatGPT, Corpus-Based Qualitative Analysis, Cultural Fidelity, José Rizal, Literary Translation.*

Introduction

Literary translation requires decisions about voice, figurative language, cultural references, and historically situated meaning. A poem may have a recognizable propositional meaning while still losing rhythm, allusion, or rhetorical force in translation. Translation theorists consequently distinguish correspondence at the lexical level from functional, interpretive, and cultural dimensions [1]–[4]. For poetry associated with anticolonial histories, lexical choices such as "homeland," "fatherland," and "country" can also carry different registers and resonances, although no single rendering is automatically authoritative [2], [3].

Recent developments in large language models (LLMs) have expanded machine translation beyond literal sentence substitution. Yet evidence on literary translation is mixed rather than uniformly negative. Karpinska and Iyyer observed that document-level LLM input improved some literary-translation judgments while critical errors persisted [5]. Wang et al. tested explanation-assisted poetry translation and found that prompting design influenced output quality [6]. Gao et al. reported strong poetry-translation results for ChatGPT relative to other systems in Chinese classical poetry, especially with form-sensitive prompts [7]. Conversely, research on literary translation evaluation shows that the choice of evaluators and metrics can substantially change rankings [8].

Research has also drawn attention to the consequences of post-editing, textual genre, and discourse context. Macken found that ChatGPT post-editing could introduce problems despite some lexical improvements [9]. Shared-task analyses suggest that literary evaluation benefits from human judgments and discourse-aware criteria [10], [11]. These studies motivate analyses that distinguish fluency from historically and aesthetically adequate interpretation rather than treating one linguistic feature as a complete quality score.

The present case focuses on selected extracts from two Spanish poems by José Rizal, "Mi Último Adiós" and "A la juventud filipina." The earlier manuscript juxtaposed passages labelled as original Spanish, human translation, and ChatGPT translation, then assigned semantic, cultural, ideological, stylistic, and metaphor codes. The original study, however, did not identify the published translator and edition for every human rendering, the complete corpus, or reproducible model-generation metadata. This revision therefore preserves the reported comparisons but bounds their interpretation. It does not present the samples as independently authenticated complete official English and Tagalog editions.

Four questions guide the comparison: (1) which semantic differences are visible in the paired examples; (2) how are culture and historical references represented; (3) how do metaphor, symbolism, tone, and form differ; and (4) which translation-shift labels best describe the documented examples? The following sections describe the corpus and coding framework, present traceable examples and qualified findings, interpret them against current scholarship, and identify data needed for replication.

Material and Methods

A. Design and Corpus

The study is a descriptive, qualitative, corpus-assisted comparison. Its source is the material presented in the supplied eight-page manuscript and appendices: selected excerpts from two poems, translations identified there as human-produced, and translations identified as ChatGPT outputs. No new text was generated to masquerade as original study data. The original manuscript describes three layers: ST (Spanish source), HT (human translation), and MT (ChatGPT output). "Corpus-based" is used in the limited sense of systematically aligned textual units, not a statistically representative corpus.

The sampling frame includes excerpts rather than full poems. The exact number of aligned units, the editions and translators of the human versions, and the original prompt logs were not provided in the source file. Accordingly, sample claims are restricted to segments reproduced in that file. The wording "official translation" is not retained without bibliographic verification. English and Tagalog comparisons should be analyzed separately in any future complete version, since target-language structure affects shifts and equivalence [1], [12].

B. Model Generation and Reproducibility

The original manuscript reports the following prompt: "Translate the following Spanish poem by José Rizal into English/Tagalog while preserving poetic meaning, tone, metaphor, and historical context." It refers to "ChatGPT-3," but this label does not identify a specific model or deployment. Date, model identifier, temperature, number of generations, access route, whether contexts were shared, and complete verbatim outputs are not documented. They must be supplied from original research records

before reproducibility can be claimed. Because LLM behavior is sensitive to prompting and contextual setup, these gaps constrain attribution [6], [7], [13].

C. Analytical Framework

The interpretation combines translation-shift description [12], thematic analysis [14], and discourse-oriented reading [15]. Text units were compared for basic semantic transfer, imagery, cultural reference, ideological register, and poetic style. Following best practice for qualitative work, codes are interpretive descriptions rather than numerical measurements or intrinsic properties of an AI system [14], [16]. The labels used in the original appendices were retained and reorganized in Table 1.

Table 1. Qualitative coding scheme retained from the source manuscript.

Domain	Illustrative codes	Operational focus
Semantics	SEM-PRES, SEM-APPROX, SEM-LOSS	Whether central propositional meaning is retained, simplified, or lost
Culture	CULT-RET, CULT-GEN, CULT-LOSS	Retention or generalization of situated references
Ideology	IDEO-PRES, IDEO-NEUT	Changes in explicit political or patriotic register
Style	STY-POET, STY-FLAT	Cadence, line structure, register, and poetic texture
Figuration	MET-PRES, MET-LOSS, SYM-RED	Metaphor or symbol maintained, recast, or omitted

Comparison proceeded in six descriptive stages reported by the original study: segmentation, open coding, code grouping, theme development, discourse interpretation, and cross-layer comparison. Unlike an empirical claim of measured agreement, the coding comparison is presented as an analytic illustration. The source appendix contains a four-row coder-agreement matrix with 100% agreement, but provides no sampling denominator, coder training protocol, or adjudication record. Its percentages cannot be independently validated. Human expert-validation claims likewise require supporting records.

D. Research Integrity and Source Verification

Poems and translation excerpts must be cross-checked against reliable source editions before publication. The short examples reproduced below follow the supplied draft; their original provenance and translator attribution are unverified. Ethical concerns primarily relate to faithful attribution, avoidance of fabricated quotations, and disclosure of generative-AI procedures. No human-participant recruitment is established in the supplied record. The authors should confirm whether any institutional review requirements apply to their own research procedures.

Results

A. Paired Text Examples

Table 2 contains two excerpts reported in the earlier document. They should be treated as documentary examples rather than a fully validated bilingual corpus. The first makes a difference in ceremonial tone visible; the second illustrates a shift from an elevated collective reference to plainer wording. The interpretations are conditional upon verification of the corresponding source and translation editions.

Table 2. Selected comparisons transcribed from the supplied manuscript.

Layer	Example 1: farewell expression	Example 2: youth expression
Spanish, as reported	Adiós, patria adorada, región del sol querida	Juventud, eres la esperanza de la patria
Human, as reported	Farewell, beloved fatherland, cherished land of the sun	Not supplied for this example
ChatGPT, as reported	Goodbye, beloved country, land of sunlight	Youth, you are the hope of the country

In the first example, "farewell" versus "goodbye" and "fatherland" versus "country" mark different levels of register. The latter can be read as simplification, but the more neutral wording is not by itself evidence that the historical message is erased. Both translations preserve a recognizable place reference. The original codebook categorized this kind of contrast as cultural generalization and stylistic flattening, an interpretation that remains contestable without full stanza context [2], [3], [12].

The second example, as reported in Appendix C of the original paper, supplies a Spanish line and a ChatGPT rendering but does not provide an aligned human version in the same displayed example. Therefore, the example cannot establish a human-versus-machine difference in the absence of a verified comparator. This editorial distinction prevents the appearance of more comparative evidence than the document actually contains.

B. Patterns Described in the Original Analysis

The original manuscript organizes its findings into metaphorical compression, cultural embeddedness, ideological neutralization, stylistic flattening, and reduction of symbolic density. The supporting passages suggest that lexical generalization and register change are plausible coding categories. However, terms such as "consistently," "significant," or "systematic" overstate the documented sample because neither a complete denominator nor a transparent frequency analysis is supplied. The findings are therefore expressed as patterns observed or proposed in the source manuscript, not prevalence estimates.

Table 3. Alignment of original research questions with defensible qualitative findings.

Research question	Observed or reported pattern	Qualification
Semantic interpretation	Examples retain broad propositional meaning with lexical substitutions	Equivalence depends on verified source text and full context
Cultural references	Some more specific words are rendered in more general terms	Generalization does not necessarily entail cultural erasure
Metaphor and poetic form	The source narrative reports figurative and stylistic simplification	Only partial aligned quotations are available
Translation shifts	A codebook identifies semantic, cultural, ideological and style categories	Frequency, prevalence, and coder reliability are not established

C. Cross-Coding Matrix

The source appendix presents agreement checkmarks for MET-LOSS, IDEO-NEUT, CULT-GEN, and STY-FLAT. As the number of segments examined per category is not described, the 100% agreement displayed there is not treated as an interrater reliability coefficient. Future research should include segment IDs, independent coding sheets, coder qualifications, and a reconciliation procedure [14], [16].

E-ISSN 2962-1674
P-ISSN 2962-5742

Discussion

A. Poetry Translation Is More Than Lexical Transfer

The reported contrasts reflect a familiar interpretive issue: linguistic fluency and propositional adequacy do not fully capture poetic function. Literary translation entails purposeful decisions about domestication, foreignization, imagery, and voice [1]–[4]. Yet a "literal" or "poetic" version cannot be judged in isolation from its intended readership and target-language conventions. The examples are most useful as prompts for close reading, not as conclusive assessments of all ChatGPT translations.

B. Context and Prompting May Change the Outcome

Recent evaluations complicate the claim that AI inherently strips poetry of metaphor or political force. Gao et al. found that ChatGPT could perform competitively in classical-poetry translation with form-focused prompts [7], while Wang et al. demonstrated that explanation-assisted strategies can improve poetry outputs [6]. Literary LLM translation quality also depends on document context [5], and even strong systems require attention to omissions and authorial voice. Consequently, the reported changes cannot be attributed solely to model architecture without controlled prompt comparisons [13], [17].

C. Human Review and Translation Quality Assessment

Expert evaluation remains important because automatic similarity scores may overlook interpretive and stylistic qualities. Zhang et al. report differences between human-evaluation protocols for literary translation [8], while automatic metric studies document the limitations of lexical and neural indicators in particular error classes [18]. Other recent work evaluates AI-assisted literary post-editing [9], creative language phenomena [19], and translation judgments elicited through explicit error-analysis prompts [20]. Together, this evidence supports a hybrid workflow in which AI drafts are reviewed by readers trained in the source language, literary form, and cultural history.

D. José Rizal, Cultural Context, and Limits of the Case

Rizal's poetry provides a historically salient setting for evaluating ideological and cultural nuance. Still, the specific meaning assigned to a line requires reliable editions and historical scholarship; it should not be inferred mechanically from a single word choice. The original dataset is small, purposively selected, and only partially documented. The source manuscript also reports a Spanish phrase "flor naciente de esperanza" without supplying a verifiable edition and line reference. This revised article omits any claim about that phrase's authenticity until the primary text is confirmed.

Several weaknesses remain methodological rather than merely editorial: the exact generated texts, sampling boundaries, human-translation bibliographies, prompt records, and expert-coding documentation are required for independent replication. Further comparative research could collect multiple outputs per model, retain complete stanza contexts, separate English and Tagalog targets, and assess poetry through a small multilingual panel. Qualitative findings should also acknowledge disconfirming examples and context-sensitive gains in AI translation [7], [8], [21], [22].

Conclusion

The selected comparisons from the supplied manuscript illustrate how translations can convey similar basic meanings while varying in register, cultural specificity, and poetic effects. These differences are appropriate topics for semantic, cultural, stylistic, and discourse-oriented coding. However, the current documentation does not support categorical claims that ChatGPT always weakens ideology or that human versions are universally superior. The study's contribution is a focused interpretive framework and an explicit set of provenance and reproducibility requirements. Future work should authenticate source and human translation editions, publish aligned text units and model-generation metadata, and use independent bilingual assessment before making generalizable comparisons.

Limitation and Suggestions

The analysis is confined to selected examples and findings reported in the supplied document. The exact corpus size, translator attribution, complete AI outputs, model version, prompt execution settings, and original coding records require author confirmation. These missing records must not be reconstructed retrospectively as though observed. Full-stanza paired datasets and expert assessment would strengthen future studies.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] L. Venuti, The Translator's Invisibility: A History of Translation, 3rd ed. London, UK: Routledge, 2017.
[2] S. Bassnett, Translation Studies, 4th ed. London, UK: Routledge, 2014.
[3] E. A. Nida, Toward a Science of Translating. Leiden, Netherlands: Brill, 1964.
[4] A. Lefevere, Translating Poetry: Seven Strategies and a Blueprint. Assen, Netherlands: Van Gorcum, 1975.
[5] M. Karpinska and M. Iyyer, "Large language models effectively leverage document-level context for literary translation, but critical errors persist," in Proc. 8th Conf. Machine Translation, 2023, pp. 419–451, doi: 10.18653/v1/2023.wmt-1.41.
[6] S. Wang, D. Wong, J. Yao, and L. Chao, "What is the best way for ChatGPT to translate poetry?," in Proc. 62nd Annual Meeting Assoc. Computational Linguistics, 2024, pp. 14025–14043, doi: 10.18653/v1/2024.acl-long.756.
[7] R. Gao, Y. Lin, N. Zhao, and Z. G. Cai, "Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator," Humanities and Social Sciences Communications, vol. 11, art. 835, 2024, doi: 10.1057/s41599-024-03363-0.
[8] R. Zhang, W. Zhao, and S. Eger, "How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs," in Proc. NAACL, 2025, doi: 10.18653/v1/2025.naacl-long.548.
[9] L. Macken, "Machine translation meets large language models: Evaluating ChatGPT's ability to automatically post-edit literary texts," in Proc. 1st Workshop Creative-text Translation and Technology, 2024, pp. 65–81. [Online]. Available: https://aclanthology.org/2024.ctt-1.7/
[10] L. Wang et al., "Findings of the WMT 2024 shared task on discourse-level literary translation," in Proc. 9th Conf. Machine Translation, 2024, pp. 699–700, doi: 10.18653/v1/2024.wmt-1.58.
[11] Y. Liu, Y. Yao, R. Zhan, Y. Lin, and D. F. Wong, "NovelTrans: System for WMT24 discourse-level literary translation," in Proc. 9th Conf. Machine Translation, 2024, pp. 980–986, doi: 10.18653/v1/2024.wmt-1.98.
[12] J. C. Catford, A Linguistic Theory of Translation. Oxford, UK: Oxford University Press, 1965.
[13] W. Jiao et al., "Is ChatGPT a good translator? A preliminary study," arXiv:2301.08745, 2023. [Online]. Available: https://arxiv.org/abs/2301.08745

[14] V. Braun and V. Clarke, "Using thematic analysis in psychology," Qualitative Research in Psychology, vol. 3, no. 2, pp. 77–101, 2006, doi: 10.1191/1478088706qp063oa.
[15] N. Fairclough, Discourse and Social Change. Cambridge, UK: Polity Press, 1992.
[16] V. Braun and V. Clarke, Thematic Analysis: A Practical Guide. London, UK: SAGE, 2022.
[17] A. Mukherjee, S. Yadav, and M. Shrivastava, "CoST of breaking the LLMs," in Proc. 9th Conf. Machine Translation, 2024, pp. 299–306, doi: 10.18653/v1/2024.wmt-1.24.
[18] T. Glushkova, C. Zerva, and A. F. T. Martins, "BLEU meets COMET: Combining lexical and neural metrics towards robust machine translation evaluation," in Proc. 24th Annual Conf. European Assoc. Machine Translation, 2023, pp. 47–58. [Online]. Available: https://aclanthology.org/2023.eamt-1.6/
[19] L. Yu, Q. Liu, and D. Xiong, "LFED: A literary fiction evaluation dataset for large language models," in Proc. LREC-COLING, 2024. [Online]. Available: https://aclanthology.org/2024.lrec-main.915/
[20] Q. Lu, B. Qiu, L. Ding, K. Zhang, T. Kocmi, and D. Tao, "Error analysis prompting enables human-like translation evaluation in large language models," in Findings ACL, 2024, pp. 8801–8816, doi: 10.18653/v1/2024.findings-acl.520.
[21] O. F. Ahmed, I. K. Dewi, and M. Y. Anis, "Translating Arabic literary metaphors: A comparative evaluation of ChatGPT-4.0 and Google Translate 2024," IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, 2025, doi: 10.24256/ideas.v13i2.7439.
[22] I. Yuliasri et al., "Student poetry translations with and without the use of AI: An evaluation of translation quality," LITE: Jurnal Bahasa, Sastra, dan Budaya, 2025, doi: 10.33633/lite.v21i2.14336.

Authors



Christopher A. Britanico is a professor at the University of Saint Anthony in Iriga City, Philippines. He holds a Master of Arts in Education and is actively engaged in academic research, scholarly writing, and research data analysis. His academic interests include higher education, adult education, and teacher education. His scholarly activities focus on contributing to educational research and advancing knowledge in teaching, learning, and educational practices across various academic contexts. (email: cbritanico@usant.edu.ph).



Nelia A. Britanico is an educator and researcher affiliated with the University of Saint Anthony, Philippines. Her academic interests include educational research, multiple intelligences, teaching and learning practices, and social and cultural studies. She has experience in teaching and supervising student research. Her scholarly contributions include studies on senior high school students' multiple intelligences and the cultural significance of tattoos in professional settings, reflecting her interdisciplinary research interests. (email: nabritanico@usant.edu.ph).



Maricon V. Quinto is an educator and researcher associated with the University of Saint Anthony in Iriga City, Philippines. Her academic interests include social studies education, historical research, cultural heritage, and community-based learning. Her research has explored the sociocultural and historical significance of local religious traditions and their implications for community life. Her scholarly interests emphasize the relationship between education, cultural identity, historical understanding, and the preservation of local heritage. (email: mvquinto@usant.edu.ph).



Marissa Del Rosario is the Dean at the University of Saint Anthony in Iriga City, Philippines. Her academic and professional interests focus on language education, teacher education, and teaching methods. She is engaged in educational research and academic leadership, with particular attention to teaching and learning practices. Her scholarly interests encompass instructional approaches, teacher development, and the advancement of educational practices that support effective teaching and meaningful learning experiences across diverse educational contexts. (email: mdrosario@usant.edu.ph).