SCRIPT JOURNAL
Journal of Linguistic and English Teaching
http://jurnal.fkip-uwgm.ac.id/index.php/Script
P-ISSN: 2477-1880; E-ISSN: 2502-6623
August 2026, Vol. 11 No. 2

Evaluating Grammarly-Assisted Post-Editing of Google Translate Outputs in Indonesian-English Academic Texts: An Exploratory Corpus Study with Expert Assessment

Ismiyati Hasiru[1], Novriyanto Napu[2], Suleman S. Bouti[3]
Universitas Negeri Gorontalo, Gorontalo, Indoensia[1,2,3]
Email Correspondence: n.napu@ung.ac.id

Abstract

Background:
Accurate Indonesian-English academic translation matters for scholarly communication, yet fluent machine output may retain grammatical, terminological, semantic, and contextual problems. This exploratory corpus study combines a descriptive profile of Grammarly flags in Google Translate (GT) output with complementary expert assessment of the processed texts.

Methodology:
An exploratory corpus-evaluation design combined descriptive analysis of Grammarly flags with complementary expert ratings and comments. The approximately 7,500-word corpus sampled published Indonesian texts from six disciplines. Six EAP instructors or professional translators evaluated processed texts using an author-developed, Machali-informed seven-point rating guide. Because the archive lacked paired segment-level and individual rater data, the analysis does not estimate pre-post improvement, flag accuracy, or inter-rater reliability.

Findings:
Of 519 Grammarly flags, correctness accounted for 285 (54.9%), clarity for 184 (35.5%), engagement for 45 (8.7%), and delivery for 5 (1.0%). These are proportions of flags, not error reductions. Expert feedback indicated that automated suggestions drew attention to surface-language features, while semantic nuance, terminology, naturalness, and cultural appropriateness required bilingual human judgment. Because individual rater data and agreement statistics were unavailable, the expert component is interpreted qualitatively rather than as a disciplinary ranking.

Conclusion:
Grammarly may serve as a monolingual screening aid for selected target-language features in GT-generated academic text, but this study does not demonstrate improved translation quality. Publication-oriented use requires expert comparison with the Indonesian source.

Originality:
The study distinguishes automated target-language feedback from bilingual translation-quality assessment in a corpus of Indonesian-English GT output.

Keywords	:	*Google Translate, Grammarly, machine translation post-editing, translation quality assessment, academic translation*
DOI	:	10.24903/sj.v11i2.2387
Received	:	May 2026
Accepted	:	August 2026
Published	:	August 2026
How to cite this article (APA)	:	Hasiru, I., Napu, N., & Bouti, S. S. (2026). Evaluating Grammarly-assisted post-editing of Google Translate outputs in Indonesian-English academic texts: An exploratory corpus study with expert assessment. *Script Journal: Journal of Linguistic and English Teaching*, *11*(2), 464-485. https://doi.org/10.24903/sj.v11i2.2387

INTRODUCTION

Machine translation (MT) is increasingly used to support cross-linguistic scholarly communication, but fluent output does not necessarily guarantee semantic accuracy or an appropriate disciplinary register. Neural MT can perform strongly on frequent patterns while remaining vulnerable to domain shift, terminology, and context-sensitive meaning (Bentivogli et al., 2016; Castilho et al., 2017; Koehn & Knowles, 2017). Recent research further shows that post-editing effort varies with text type and that source-language competence remains important to post-editing performance (Lesznyák et al., 2024; Wang & Daghigh, 2024). Generative translation and post-editing can also produce hallucinated or untrustworthy edits (Guerreiro et al., 2023; Raunak et al., 2023). Post-editing is therefore required when MT output is intended for consequential academic communication (Nitzke & Hansen-Schirra, 2021).

Grammarly is an automated writing feedback tool designed mainly to identify target-language patterns involving grammar, mechanics, clarity, and style. Direct accuracy research has identified overflagging, false positives, and optional-use recommendations in published academic prose, while a learner study found that participants correctly addressed 85% of Grammarly-flagged usages and that feedback accuracy influenced response accuracy (Abu Qub'a et al., 2024; Guo et al., 2022). Broader classroom and review evidence suggests potential support for surface-level revision but also overcorrection, context-insensitive suggestions, and limited higher-order feedback (Barrot, 2023; Dizon & Gayed, 2024; Koltovskaia, 2023; Ranalli & Yamashita, 2022).

The substantive unresolved problem is the absence of evidence that a monolingual writing assistant's proprietary flags correspond to source-aware machine-translation evaluation constructs such as adequacy, meaning preservation, and terminological accuracy. Open-access MT evaluation benchmarks use aligned source-target material, explicit error criteria, qualified evaluators, and document context (Freitag et al., 2021; Kocmi & Federmann, 2023), whereas Grammarly research predominantly examines English-language writing feedback. The present study therefore separates Grammarly flag behavior from source-aware translation-quality judgment in Indonesian-English academic texts. It profiles Grammarly flags in GT output and examines expert concerns about the processed texts. It does not evaluate classroom instruction, student learning, or writing development; English for Academic Purposes (EAP) is treated only as the academic communication context.

Academic texts require explicit argumentation, precise terminology, coherent information structure, and a register appropriate to disciplinary readers. These requirements make academic translation particularly sensitive to formulations that are grammatically plausible but semantically or pragmatically inaccurate. Automated writing feedback can identify some target-language patterns, but its pedagogical value and revision quality depend on how users interpret and verify its suggestions (Godwin-Jones, 2022; Zhang & Hyland, 2018).

For Indonesian scholars working across languages, MT and automated feedback tools may offer practical assistance. However, the present corpus-based study cannot establish learning gains or instructional effectiveness. Any educational use of the GT-Grammarly workflow is therefore considered a prospective application that requires evaluation with students, instructional procedures, and learning outcomes.

Translation requires the transfer of meaning from a source text into a target text while respecting target-language norms and the communicative purpose of the text (Bassnett, 2002). Contemporary neural and generative MT models have substantially improved fluency, yet domain-specific terminology, long-range context, semantic adequacy, and hallucination remain persistent challenges (Guerreiro et al., 2023; Koehn & Knowles, 2017; Vaswani et al., 2017; Wu et al., 2016). Professional translators working with source text and document context may produce materially different quality judgments from monolingual or crowd evaluations (Freitag et al., 2021). Although large-language-model evaluators offer promising automated assessment, their performance is benchmarked against human quality labels and does not eliminate the need for source-aware validation (Kocmi & Federmann, 2023). Consequently, fluent MT output should not be equated with an accurate translation.

Grammarly provides automated feedback on target-language form, including grammar, punctuation, concision, delivery, and vocabulary. Because the tool operates primarily on the English target text, it cannot independently determine whether a revision preserves the meaning of the Indonesian source. Grammarly is therefore conceptualized here as a monolingual post-editing aid rather than as a translation-quality evaluator. Previous research likewise indicates that automated feedback is most defensible when users critically evaluate suggestions instead of accepting them automatically (Dembsey, 2017; Koltovskaia, 2020; Ranalli & Yamashita, 2022).

Translation Quality Assessment (TQA) concerns accuracy, acceptability, and communicative effectiveness. Machali's (2009) framework directs attention to linguistic,

semantic, pragmatic, terminological, and stylistic aspects of translation. In this study, these dimensions informed an author-developed seven-point rating guide rather than a validated Machali scale; the operational mapping and validation limitations are reported in the Methodology. Grammarly's proprietary categories and Machali-informed expert judgments were treated as distinct evidence sources: the former described automated flags, whereas the latter captured human evaluations of translation quality. This distinction prevents automated suggestions from being treated as confirmed errors.

Against this background, the study addressed three research questions: (1) What types and proportions of Grammarly flags occurred in GT-translated academic texts? (2) How did these patterns vary descriptively across the six disciplinary samples, subject to corpus-size constraints? (3) What qualitative concerns about semantic adequacy, terminology, naturalness, readability, and publication suitability were identified in the retained expert comments? These questions define an exploratory corpus evaluation rather than an educational intervention.

METHODOLOGY

Research Design

This study adopted an exploratory corpus-evaluation design in which quantitative profiling of Grammarly flags had descriptive priority and expert ratings and open-ended comments provided complementary assessment. This designation is preferable to a sequential explanatory mixed-method design because the archived record does not show that quantitative results prospectively determined rater selection, qualitative sampling, or interview prompts. Both forms of evidence were collected within the April-June 2024 observation period and were integrated only during interpretation. This choice follows the principle that design labels should match the actual sequence and point of integration (Creswell & Plano Clark, 2018; Fetters et al., 2013; Napu & Manipuspika, 2025). The study was informed by Machali's (2009) TQA dimensions rather than treating Grammarly categories as a translation-quality framework. A purposive corpus of approximately 7,500 words was selected from peer-reviewed Indonesian-language articles in the Neliti database and represented linguistics, technology, economics, engineering, medical science, and law. Purposive selection was appropriate for assembling information-rich texts across contrasting disciplinary contexts rather than estimating population prevalence (Palinkas et al., 2015). The source record indicated 10-12 articles per discipline; however, discipline-specific word totals were not available in the manuscript archive. Accordingly, disciplinary counts are interpreted descriptively and are not presented as normalized error rates.

The two evidence sources were integrated during interpretation by comparing the frequency patterns of Grammarly flags with rater concerns involving accuracy, naturalness, readability, terminology, and cultural appropriateness. They answered related but non-equivalent questions: the flag counts described what the tool marked, whereas the expert comments identified translation-quality concerns requiring human judgment. No connecting phase used the flag results to determine qualitative sampling or prompts; therefore, the expert component is described as complementary assessment rather than an explanatory qualitative strand. This interpretive comparison did not constitute a paired intervention-effect analysis.

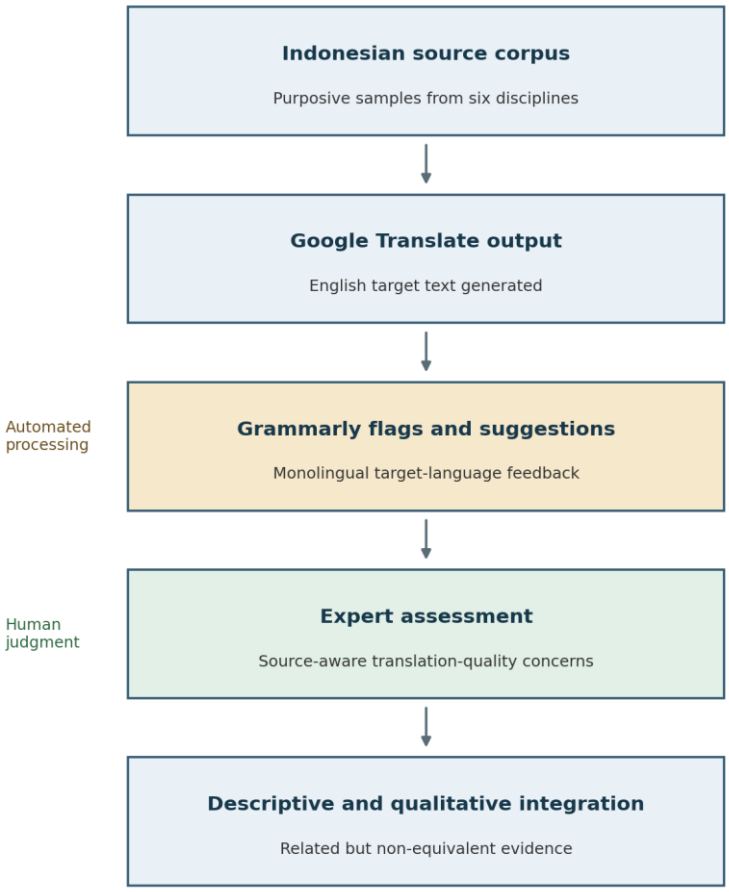

Figure 1. Methodological workflow and integration of evidence sources.

Participants and Data Sources

Corpus construction is interpreted in terms of disciplinary breadth rather than statistical representativeness. Together, the six fields expose the workflow to different lexical, syntactic, and rhetorical demands: linguistics and law contain terminology and argument structures that are highly context dependent; engineering and technology frequently employ procedural and technical formulations; economics uses quantitative and institutional language; and medical

science requires precise terminology and cautious claims. The corpus should therefore be understood as a purposive cross-domain sample for exploratory evaluation. It was not designed to estimate the prevalence of translation problems in Indonesian scholarship as a whole. Because the archive preserves only the aggregate corpus length and not the number of words contributed by each discipline, the study cannot determine whether a larger disciplinary count reflects a higher flag density, a larger amount of text, longer sentences, or a different distribution of document types.

The unit of quantitative analysis was an individual flag generated by Grammarly. A flag is a platform-defined notification attached to a span of English text and may concern correctness, clarity, engagement, or delivery. This unit differs from a linguistically adjudicated error: one sentence may produce several overlapping flags, while a serious source-target meaning problem may produce no flag at all. Frequencies were therefore counted as instances of tool behavior rather than independent observations of translation failure. The unit of qualitative analysis was a meaning-bearing statement in the retained rater comments. Statements were grouped according to the aspect of translation quality they addressed, including accuracy, naturalness, readability, terminology, sentence structure, and cultural or contextual appropriateness. This use of coded meaning-bearing segments to identify recurring patterns is consistent with thematic analysis principles (Braun & Clarke, 2006). This distinction between units allowed the quantitative strand to characterize what the tool noticed and the qualitative strand to characterize what expert readers considered consequential.

Instruments and Data Collection

The processing sequence was operationalized as a one-directional workflow between April and June 2024. Indonesian passages were first entered into the web-based GT interface; the resulting English text was then transferred to Grammarly; and the categories, flagged spans, and suggested revisions visible during the observation period were recorded. The analysis did not assume that every suggestion was accepted, nor did it equate the presence of a suggestion with a beneficial edit. Instead, each recorded item represented a candidate revision requiring contextual evaluation. This decision is important because Grammarly analyzes the target language without direct access to the Indonesian source. It can identify patterns associated with English usage, but it cannot independently establish whether a suggested deletion, substitution, or restructuring preserves propositional meaning, legal qualification, disciplinary terminology, or culture-specific reference.

Expert evaluation was used to extend the analysis beyond the proprietary flag taxonomy. Eligibility required advanced English proficiency and relevant experience in at least one of EAP instruction, professional translation, or academic editing. The archived records identified six eligible raters but did not preserve each person's disciplinary field, years of experience, or individual score matrix; no additional qualifications are inferred. The author-developed seven-point guide was informed by Machali's dimensions. Linguistic form was operationalized as grammar and syntax; semantic adequacy as preservation of source meaning; pragmatic appropriateness as communicative fit; terminology as accuracy of domain-specific wording; and stylistic suitability as naturalness, readability, and academic register. These dimensions functioned as complementary criteria rather than as a mechanically summed scale. No surviving documentation established formal external review or pilot testing of the guide; it is therefore not described as a validated instrument. Contemporary MT evaluation research supports the use of qualified translators, aligned source-target material, document context, and explicit error criteria (Freitag et al., 2021). The retained disciplinary means summarize the available evaluations, but the absence of individual score matrices prevents reconstruction of dispersion, internal consistency, or inter-rater agreement. The open-ended comments were consequently used to explain recurring concerns rather than to calculate theme prevalence. No claim is made that the qualitative categories are exhaustive or that thematic saturation was achieved. This conservative treatment preserves the evidentiary value of the experts' observations while acknowledging the limits of the archived rating and coding record.

Data Analysis

Descriptive statistics were selected because the available data do not satisfy the requirements for inferential comparison. Category frequencies and percentages describe the composition of all 519 flags. Overall flag density was calculated as the number of flags divided by the approximately 7,500-word corpus and multiplied by 1,000. Because the exact corpus word count was not retained, rates are reported as rounded approximations and support only corpus-level interpretation. Equivalent disciplinary rates would require the word total for each field, and valid paired tests would require matched raw GT and Grammarly-processed segments scored with the same criteria. Neither dataset was retained. Accordingly, the study reports overall density, category and subcategory frequencies, selected descriptive means, and qualitative patterns without significance tests, effect sizes, confidence intervals, or causal language. This analytic boundary is consistent with the exploratory purpose and prevents apparent numerical precision from exceeding the information contained in the source record.

After data collection, the recorded flags were tabulated by category and subcategory, converted to shares of all flags, and expressed as approximate corpus-level rates per 1,000 words. The retained expert comments were then interpreted alongside those frequency patterns; discipline-level means were retained only as archival descriptive information and were not used to rank fields. Direct Grammarly-accuracy studies have used human review to distinguish confirmed errors, false positives, and acceptable alternatives (Abu Qub'a et al., 2024; Guo et al., 2022). Because the present archive lacked the aligned source passages, raw GT outputs, exact flags and suggestions, and individual adjudications required to apply that benchmark retrospectively, no precision, false-positive rate, or stylistic-alternative rate is reported. The analysis likewise does not report dispersion or inter-rater reliability or estimate improvement attributable to Grammarly.

FINDINGS AND DISCUSSIONS

Grammarly Flags in GT-Translated Texts

The quantitative analysis identified 519 Grammarly flags in GT-translated text. Correctness accounted for 285 flags (54.9%), clarity for 184 (35.5%), engagement for 45 (8.7%), and delivery for 5 (1.0%). Relative to the approximately 7,500-word corpus, these categories correspond to about 38, 25, 6, and 1 flag per 1,000 words, respectively, or approximately 69 flags per 1,000 words overall. These values describe the distribution and approximate density of automated flags; they do not represent reductions in errors or demonstrated improvements in translation quality. Accordingly, no precision, false-positive rate, or stylistic-alternative rate can be calculated for this corpus. This boundary is important because direct accuracy studies have documented both overflagging and acceptable optional-use cases (Abu Qub'a et al., 2024; Guo et al., 2022).

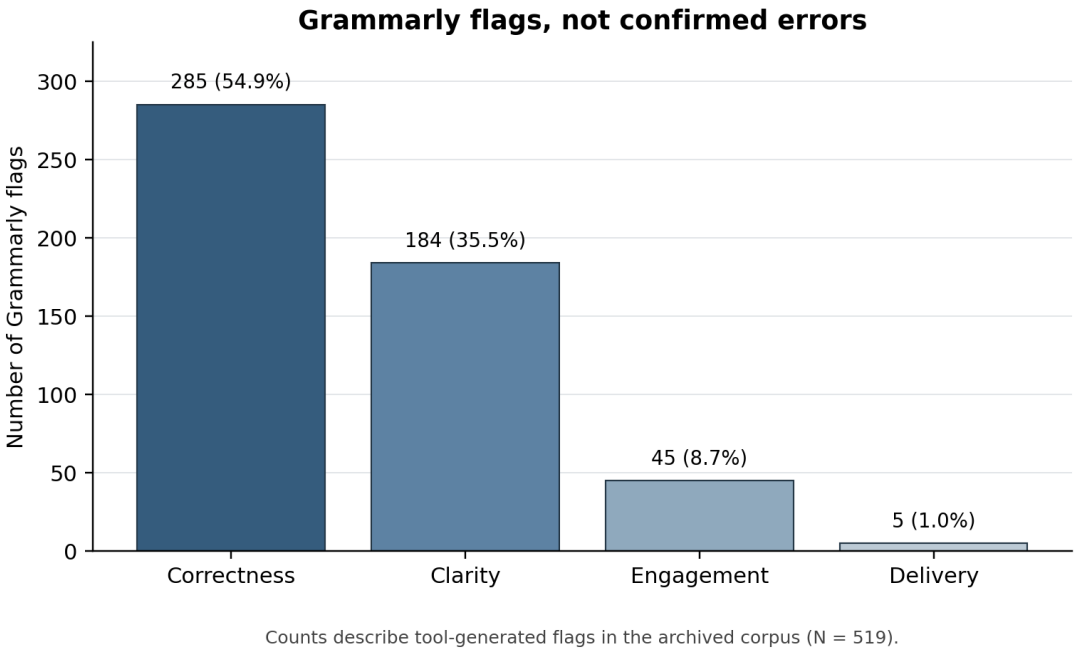

Figure 2. Distribution of Grammarly flags across the four platform-defined categories.

The flags were further divided into subcategories. Because the flags were not independently adjudicated and discipline-specific word denominators were unavailable, the following results are presented as tool-generated indicators requiring human review rather than as confirmed errors or valid cross-disciplinary error rates.

Table 1 consolidates representative target-language examples. Because the Indonesian source segments were not retained in the archived dataset, the first column reports their raw Google Translate output; Grammarly suggestions and the judgments are therefore limited to target-language implications.

Table 1. Representative Grammarly feedback records and bounded editorial judgments

Raw GT output	Grammarly flag/suggestion	Editor/expert judgment
…there is always a feeling of a lack of space in the warehouse.	Delete ‘a feeling of’ for concision.	May improve concision; source comparison is required to verify preserved emphasis.
The results of the classification are influenced by the segmentation process…	Rewrite in active voice.	Shorter wording changes thematic emphasis; accept only after source-aware review.
this is evidenced by the average value of 5.75 from a scale of 1 to 7	Replace ‘from’ with ‘on’.	The suggestion improves English idiomaticity but does not establish translation adequacy.
empon-empon	Replace with ‘upon’ or add to dictionary.	Replacement would distort a culture-specific term; retain or explain after source verification.

Raw GT output	Grammarly flag/suggestion	Editor/expert judgment
the recent situation	Replace 'recent' with 'current'.	This is a contextual collocational preference, not a confirmed translation error.

Clarity

The clarity category included flags for concision, passive voice, readability, and unclear sentences. Concision accounted for 113 flags across the six disciplines. The following examples reproduce GT output and Grammarly suggestions; because aligned Indonesian source segments were not retained in the report, they illustrate target-language feedback but cannot establish semantic accuracy.

Other problems encountered are finding materials that take a long time, the total distance for moving is not optimal, and there is always a feeling of a lack of space in the warehouse.

It can be seen that Grammarly flagged "a feeling of" as an empty phrase and suggested its deletion. The suggestion may improve target-language concision, but it cannot be classified as a translation improvement without comparison with the Indonesian source and the intended emphasis.

Grammarly also produced two passive-voice flags in the linguistic and technology samples and recommended active-voice revisions. Passive voice is not inherently erroneous in academic prose; its appropriateness depends on disciplinary convention, discourse function, and intended emphasis.

The results of the classification are influenced by the segmentation process carried out and the input parameters used during the training process.

It shows that Grammarly suggested the active-voice version, "The segmentation process influences the result of classification." Although this version is shorter, it shifts thematic emphasis from the classification results to the segmentation process. Expert review is therefore necessary to determine whether the change preserves source-text meaning and academic information structure.

Readability issues were identified in economics, technology, law, and medical texts, with a total of seven issues. For example:

If a plot of land is used as a dowry, it means that it must comply with Indonesian Land Law, namely Law (UU) Number 5 of 1960 concerning Basic Agrarian Regulations, commonly called the Basic Agrarian Law (UUPA), along with various implementing regulations, especially regarding the legal substance governing the transfer of land rights that requires

registration according to the legal norms contained in Government Regulation (PP) Number 24 of 1997 concerning Land Registration.

Grammarly suggested dividing this long sentence into two sentences. The recommendation addresses target-language readability, but the absence of an aligned source passage and segment-level expert annotation prevents confirmation that the proposed restructuring preserves every legal relation and qualification.

Grammarly suggested revisions to simplify sentence structures, eliminate jargon, and improve paragraph length. Unclear sentence issues were detected in all subject categories, totaling 62 issues. For example:

This study took a sample of 52 companies that carried out corporate rights issues in 2018 and 2019 which were listed on the Indonesia Stock Exchange, using purposive sampling

It can be seen that Grammarly suggested fronting the sampling phrase: "Using purposive sampling, this study took a sample of 52 companies that carried out corporate rights issues in 2018 and 2019, which were listed on the Indonesia Stock Exchange." This may improve information flow, but the relative-clause attachment remains potentially ambiguous and requires human editing.

Collectively, the clarity flags identify passages that warrant review for concision and readability. They do not, by themselves, demonstrate that GT was inaccurate or that accepting every suggestion would improve translation quality.

Correctness

Correctness was Grammarly's largest category, with 285 of 519 flags in the archived category total. It included platform-defined subcategories for consistency, conventions, grammar, punctuation, and spelling. These labels describe the tool's feedback taxonomy rather than an independently validated error classification. The recoverable named subcategories account for 284 flags (2 consistency, 8 conventions, 69 grammar, 166 punctuation, and 39 spelling); one archived correctness flag could not be reconciled with a named subcategory and is therefore reported as unclassified rather than assigned without evidence.

using a protocol with a composition almost the same as what is currently used in RSKD, indicating that the effectiveness of this protocol is approximately 70%. of children with ALL

It shows that Grammarly suggested capitalizing "using" because it begins the sentence. The fragmentary context, however, limits any conclusion about the source of the problem or the adequacy of the full translation.

Grammarly flagged capitalization and punctuation patterns in the economics and medical samples. Grammar-related flags occurred 69 times across the corpus, particularly in economics, linguistics, and technology. For example:

this is evidenced by the average value of 5.75 from a scale of 1 to 7

Grammarly flagged the preposition "from" and suggested "on," producing the idiomatic expression "on a scale of 1 to 7."

This example concerns target-language prepositional usage. It should not be generalized as evidence that GT systematically fails with specialized terminology because the study did not compare error rates with another translation system or independently validate each Grammarly flag.

Punctuation-related issues were the most prevalent, with 166 instances detected across all fields. For example:

The representation in question is AB's way of describing the social situation in DKI Jakarta which includes…

Grammarly recommended a comma before "which." Whether the clause is restrictive or nonrestrictive must be determined from the intended meaning.

Punctuation flags occurred across disciplines. Their distribution may reflect corpus composition, sentence length, source-text punctuation, or Grammarly's detection sensitivity rather than a general defect in GT. Punctuation suggestions can improve readability, but their relevance to translation quality must be judged against the source text and disciplinary convention.

Spelling

Grammarly generated 39 spelling flags across the corpus; no spelling flags were recorded in the linguistic sample. Engineering, law, and medical texts produced the largest counts, often because Indonesian or culture-specific terms were absent from Grammarly's lexicon. For example,

empon-empon

The word "empon" in this text refers to Javanese traditional drink, made from spices that are beneficial for health. Grammarly suggested to change this to "upon" or add the word to the personal dictionary so it will not be marked as misspelled in the future.

Grammarly treated untranslated or culturally specific forms as possible misspellings because they were absent from its lexicon. The "empon-empon" example demonstrates a false-positive risk: replacing the term with "upon" would alter the intended meaning. Domain

knowledge, source-text consultation, and appropriate explanation or retention of culture-specific terminology are therefore required.

Delivery

Delivery flags concerned formality, tone, and presentation. Only five delivery flags (1.0% of all flags) were recorded. This low count indicates that Grammarly seldom activated this category in the corpus; it does not establish that GT reliably preserved formality or disciplinary register.

important parameter in measuring the quality of information systems and/or software

Grammarly flagged "and/or" as potentially unsuitable for formal prose and suggested selecting one conjunction or rewriting the phrase. The correct revision depends on the intended logical relation.

Grammarly can draw attention to formulations such as "and/or" or sentence fragments, but the appropriateness of a revision depends on the intended logical relation and disciplinary convention. Delivery flags are therefore best interpreted as prompts for editorial judgment rather than automatic corrections.

Engagement

The engagement category contained 45 flags, including 42 vocabulary-related flags and three fluency flags. The flags were distributed across disciplines, but valid disciplinary rate comparisons could not be calculated without field-specific word denominators.

the identity of DKI Jakarta in the past, the recent situation…

Grammarly detected unusual word pair in the sample sentence "recent situation" and suggested to change it to "current situation"

The suggested change from "recent situation" to "current situation" reflects a collocational preference rather than a demonstrated translation error. Either expression may be appropriate depending on temporal reference and source-text meaning, illustrating the need for contextual validation.

One of the key risk groups is the injecting drug user group

Grammarly detected overused vocabulary in the sample sentence and suggested to change the word "key" to "vital"

Similarly, replacing "key" with "vital" is not necessarily more precise or more academic; the preferred word depends on context and intended emphasis. Engagement flags should therefore be interpreted as optional stylistic prompts, not evidence of translation inaccuracy.

Across categories, Grammarly flags were concentrated in correctness and clarity. The evidence supports a limited conclusion: Grammarly frequently identifies target-language features that warrant human review. It does not establish causal improvement because raw GT output and Grammarly-processed output were not evaluated using the same paired quality criteria (Fiederer & O'Brien, 2009; Nitzke & Hansen-Schirra, 2021).

The concentration of flags in correctness and clarity is partly consistent with Grammarly's design as a target-language writing assistant. These categories contain features that can often be approximated from local linguistic patterns, such as punctuation, prepositions, sentence length, and lexical combinations. Their prominence should not be read as a hierarchy of translation problems. Semantic omission, mistranslation, altered scope, and inappropriate disciplinary terminology may be less visible to a monolingual system even when they are more consequential for readers. Conversely, a stylistic flag may identify an acceptable construction and thus function as an invitation to review rather than a diagnosis. Research on automated written corrective feedback similarly shows that system feedback can be useful but imperfect, making user evaluation and feedback literacy central to responsible uptake (Koltovskaia, 2023; Ranalli & Yamashita, 2022). The present distribution is therefore best interpreted as a profile of the tool's attention within this corpus, not as a complete model of translation quality.

Raters' Qualitative Translation-Quality Concerns

Machali-informed open-ended comments provided human evidence complementary to the automated flag counts. The defensible concerns involved unnatural or stiff phrasing, imprecise vocabulary, problematic sentence structure, unresolved semantic meaning, culture-specific terminology, and insufficient academic register. These qualitative observations indicate that surface-level recommendations did not consistently resolve source-dependent or publication-oriented problems. Because standard deviations, individual-level scores, and inter-rater reliability were unavailable, the archived means are not used to rank disciplines. The evidence therefore supports a bounded role for Grammarly as a target-language screening aid.

The archived discipline-level means are retained only to document the original evaluation record. They are not interpreted as evidence that one discipline is easier to translate or that the workflow performs more reliably in engineering or technology than in linguistics. Without normalized disciplinary denominators, individual ratings, dispersion estimates, reliability statistics, and matched source-target annotations, the available evidence can defend only the qualitative concerns reported above. Domain sensitivity remains a question for a larger stratified study.

Translation Difficulties

One rater stated, "Diperlukan pembenahan terhadap terjemahan yang ada agar lebih layak terbit di publikasi" (The existing translations require revision to become more suitable for publication). This comment indicates that at least one expert did not regard the processed output as publication-ready. It is reported as an individual qualitative judgment and is not generalized to every rater or corpus segment. The observation nevertheless reinforces the distinction between surface-language feedback and comprehensive bilingual quality assessment.

GT and Grammarly in Translation Workflow

Grammarly operates on the English target text and cannot directly evaluate source-target fidelity. Its grammar, punctuation, spelling, clarity, and style flags therefore represent monolingual feedback rather than bilingual translation validation (Nitzke & Hansen-Schirra, 2021). This distinction is consistent with evidence that source-language reading and document context materially affect post-editing and expert MT evaluation (Freitag et al., 2021; Lesznyák et al., 2024).

Rater comments indicated that Grammarly drew attention to some surface features but did not resolve all contextual or semantic concerns. Because paired GT-versus-Grammarly quality ratings were unavailable, this observation is interpreted qualitatively rather than as measured improvement.

Literal source-language structure and culture-specific meaning require comparison of the Indonesian source with the English target. These dimensions exceed a monolingual writing assistant's evidentiary scope and require bilingual expert judgment.

Raters described some translations as stiff or insufficiently academic in tone, suggesting that target-language fluency and disciplinary register were not fully captured by automated feedback. These perceptions should be interpreted cautiously because dispersion and agreement statistics were unavailable.

Grammarly may assist with reviewing selected language-mechanics features, but expert translators remain necessary for meaning, context, terminology, cultural nuance, and disciplinary style. Automated feedback should therefore complement rather than replace bilingual post-editing.

Grammarly as a Monolingual Post Editor

In this study, Grammarly is best understood as an automatic monolingual post-editing aid. Monolingual post-editing revises a machine-generated target text without necessarily

consulting the source and therefore primarily addresses target-language fluency and comprehensibility rather than source-text accuracy (Nitzke & Hansen-Schirra, 2021).

This conceptualization explains why grammatical or readability suggestions may coexist with unresolved semantic and cultural problems. Grammarly can identify candidate revisions in English, but it cannot confirm whether they preserve the meaning and function of the Indonesian source.

The raters' observations of unnaturalness, stiffness, and culture-related problems are consistent with this limitation. A monolingual system may produce a well-formed English sentence while overlooking mistranslation, omission, addition, or a contextually inappropriate lexical choice.

Accordingly, erroneous or misleading MT output may remain undetected when a post-editor does not compare the target text with the source.

Grammarly's contribution should therefore be evaluated at the level of target-language assistance, not as independent evidence of translation accuracy. High-quality post-editing requires source-text consultation and informed human decision-making (Fiederer & O'Brien, 2009; Nitzke & Hansen-Schirra, 2021).

This interpretation is consistent with translation research showing that greater target-language fluency does not necessarily guarantee greater adequacy and that human evaluation remains necessary for meaning-sensitive judgments (Castilho et al., 2017; Koehn & Knowles, 2017).

The findings motivate, but do not test, a staged account of human-AI collaboration. As a proposed implication, automated assistance could be used to locate candidate surface-level problems, after which bilingual experts would decide whether a change is necessary, preserves source meaning, and conforms to disciplinary expectations. The present study did not measure time on task, keystrokes, cognitive effort, or workflow efficiency; therefore, it cannot establish reduced effort or productivity. Such claims require process-oriented evidence of the kind used in eye-tracking and keyboard-logging research on post-editing (Wang & Daghigh, 2024). The risk is not limited to missed errors. Users may accept fluent suggestions that weaken precision, remove necessary emphasis, or replace a valid culture-specific term. Recent work on large-language-model post-editing likewise reports improvements alongside hallucinated or untrustworthy edits, demonstrating that stronger generative capacity does not eliminate the need for human validation (Raunak et al., 2023). Research on automated feedback for scientific writing also distinguishes surface revision from substantive assessment and emphasizes the

value of specific, actionable feedback grounded in the manuscript's content (Chamoun et al., 2024). In the present context, a proposed workflow would separate automatic detection, bilingual verification, disciplinary review, and final authorial approval rather than treating a Grammarly suggestion as the endpoint of post-editing.

User Expertise and Interpretation of Automated Feedback

Automated feedback is not self-validating. Users need sufficient linguistic and disciplinary competence to accept, modify, or reject suggestions, especially for terminology, collocation, register, and meaning-sensitive choices.

Koltovskaia (2020) found that engagement with Grammarly feedback varies and that users do not necessarily interpret or implement automated suggestions consistently. Studies of Grammarly used alongside human feedback similarly support a complementary rather than substitutive role for the tool (Koltovskaia, 2023; O'Neill & Russell, 2019).

In the present corpus, the "empon-empon," "recent situation," and "key/vital" examples illustrate why automated suggestions require contextual judgment. A mechanically accepted recommendation can be unnecessary, stylistically debatable, or semantically harmful.

Prior classroom research conceptualizes cognitive engagement as learners' interpretation, evaluation, and use of feedback during revision (Zhang & Hyland, 2018). However, the present study did not examine students, classroom interaction, or revision behavior. Cognitive engagement is therefore a relevant implication for future EAP research, not an outcome demonstrated by this corpus analysis.

The defensible conclusion is that Grammarly can supply rapid candidate feedback, while the quality of any resulting revision depends on human verification against linguistic context, disciplinary expectations, and the source text.

Implications for Tool Development and Future Research

The observed false-positive and context-sensitive examples suggest that future writing-assistance systems could benefit from domain-aware lexicons, terminology controls, and interfaces that display source and target segments together. These are design implications derived from the corpus observations, not evaluated outcomes. Future studies should archive aligned Indonesian source, raw GT output, Grammarly suggestion, final revision, and expert rationale; compare versions with identical criteria; and test educational applications longitudinally in actual EAP classrooms.

A stronger future evaluation should prospectively archive every stage of this sequence. For each sampled segment, the dataset should contain the Indonesian source, unedited GT

output, the exact Grammarly flag and suggestion, the editor's accept-or-reject decision, the final version, and a bilingual rationale. Independent experts should first adjudicate whether each flag is a true positive, false positive, or stylistic alternative and then score both the raw and edited outputs with identical, predefined criteria. Reporting should include corpus words by discipline, issue rates per 1,000 words, means and standard deviations, agreement coefficients, confidence intervals, and effect sizes where assumptions are satisfied. Such a design would permit separate estimates of detection precision, revision benefit, semantic harm, and disciplinary variation. If the workflow is subsequently tested in EAP education, student proficiency, feedback literacy, time on task, revision behavior, and delayed learning outcomes should be measured independently of text-level quality. These procedures would transform the present exploratory observations into a reproducible test of both technological and pedagogical effectiveness.

Research Limitations

The study is limited by the small corpus; unavailable discipline-specific word totals; absence of independently adjudicated, aligned source-to-revision records; and unavailable tool-build, settings, and individual-rater information. These constraints precluded normalized disciplinary comparisons, paired pre-post quality estimates, flag-accuracy measures, dispersion statistics, and inter-rater reliability. The findings are limited to GT and Grammarly and should therefore be interpreted as an exploratory description of automated flags and retained expert perceptions, not as evidence of learning, efficiency, publication readiness, or generalizable translation improvement.

CONCLUSION

This study provides a corpus-based description of Grammarly flags in GT-generated Indonesian-English academic texts and of expert perceptions of the processed output. The flags were concentrated in correctness and clarity, while expert feedback identified unresolved semantic, contextual, terminological, naturalness, and register-related concerns. Grammarly should therefore be treated as a monolingual screening aid whose suggestions remain provisional until verified against the Indonesian source and disciplinary norms. The principal limitation is the absence of aligned source-to-revision data and independently adjudicated, individual-level ratings. Future studies should archive each processing stage and apply blinded bilingual evaluation to test detection accuracy, revision benefit, semantic harm, efficiency, and educational outcomes.

DECLARATION OF GENERATIVE AI

The authors acknowledge the use of ChatGPT (OpenAI) to assist in the preparation of this manuscript. ChatGPT was used to assist with summarizing selected literature, integrating relevant sources and findings of other scholars, providing definitions and explanations of key terms and concepts, and improving the clarity and readability of the manuscript. The outputs from these prompts were used as supporting assistance during the writing process. The authors reviewed and evaluated the outputs and made the final decisions regarding their accuracy, relevance, and incorporation into the manuscript. While the authors acknowledge the use of AI, the authors remain the sole author of this article and take full responsibility for the content therein, including its arguments, analysis, interpretations, conclusions, and the appropriate use and presentation of sources.

REFERENCES

Abu Qub'a, A., Abu Guba, M. N., & Fareh, S. (2024). Exploring the use of Grammarly in assessing English academic writing. Heliyon, 10(15), Article e34893. https://doi.org/10.1016/j.heliyon.2024.e34893

Barrot, J. S. (2023). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. Computer Assisted Language Learning, 36(4), 584-607. https://doi.org/10.1080/09588221.2021.1936071

Bassnett, S. (2002). Translation studies (3rd ed.). Routledge. https://doi.org/10.4324/9780203427460

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: A case study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 257-267). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1025

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art? The Prague Bulletin of Mathematical Linguistics, 108, 109-120. https://doi.org/10.1515/pralin-2017-0013

Chamoun, E., Schlichtkrull, M., & Vlachos, A. (2024). Automated focused feedback generation for scientific writing assistance. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 9742-9763). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.580

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE Publications. https://us.sagepub.com/en-us/nam/product/978-1483344379

Dembsey, J. M. (2017). Closing the Grammarly gaps: A study of claims and feedback from an online grammar program. The Writing Center Journal, 36(1), 63-100. https://www.jstor.org/stable/44252638

Dizon, G., & Gayed, J. M. (2024). A systematic review of Grammarly in L2 English writing contexts. Cogent Education, 11(1), Article 2397882. https://doi.org/10.1080/2331186X.2024.2397882

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs: Principles and practices. Health Services Research, 48(6 Pt 2), 2134-2156. https://doi.org/10.1111/1475-6773.12117

Fiederer, R., & O'Brien, S. (2009). Quality and machine translation: A realistic objective? The Journal of Specialised Translation, 11, 52-74. https://www.jostrans.org/issue11/art_fiederer_obrien.php

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics, 9, 1460-1474. https://doi.org/10.1162/tacl_a_00437

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. Language Learning & Technology, 26(2), 5-24. https://doi.org/10.64152/10125/73474

Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., & Martins, A. F. T. (2023). Hallucinations in large multilingual translation models. Transactions of the Association for Computational Linguistics, 11, 1500-1517. https://doi.org/10.1162/tacl_a_00615

Guo, Q., Feng, R., & Hua, Y. (2022). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? Computer Assisted Language Learning, 35(9), 2312-2331. https://doi.org/10.1080/09588221.2021.1879161

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation (pp. 193-203). European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.19/

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation (pp. 28-39). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3204

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback provided by Grammarly: A multiple case study. Assessing Writing, 44, Article 100450. https://doi.org/10.1016/j.asw.2020.100450

Koltovskaia, S. (2023). Postsecondary L2 writing teachers' use and perceptions of Grammarly as a complement to their feedback. ReCALL, 35(3), 290-304. https://doi.org/10.1017/S0958344022000179

Lesznyák, M., Bakti, M., & Sermann, E. (2024). Reading takes it all? The role of language competence and subject knowledge in legal translation. The Interpreter and Translator Trainer, 18(2), 192-211. https://doi.org/10.1080/1750399X.2024.2344177

Machali, R. (2009). Pedoman bagi penerjemah: Panduan lengkap bagi Anda yang ingin menjadi penerjemah profesional. Kaifa. https://opac.perpusnas.go.id/DetailOpac.aspx?id=151632

Napu, N., & Manipuspika, Y. S. (2025). Pengantar metodologi penelitian dalam kajian penerjemahan. Deepublish. https://deepublishstore.com/produk/buku-pengantar-metodologi-penelitian-dalam-kajian-penerjemahan/

Nitzke, J., & Hansen-Schirra, S. (2021). A short guide to post-editing. Language Science Press. https://doi.org/10.5281/zenodo.5646896

O'Neill, R., & Russell, A. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. Australasian Journal of Educational Technology, 35(1), 42-56. https://doi.org/10.14742/ajet.3795

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Administration and Policy in Mental Health and Mental Health Services Research, 42, 533-544. https://doi.org/10.1007/s10488-013-0528-y

Ranalli, J., & Yamashita, T. (2022). Automated written corrective feedback: Error-correction performance and timing of delivery. Language Learning & Technology, 26(1), 1-25. https://doi.org/10.64152/10125/73465

Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., & Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 12009-12024). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.804

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762

Wang, Y., & Daghigh, A. J. (2024). Effect of text type on translation effort in human translation and neural machine translation post-editing processes: Evidence from eye-tracking and keyboard-logging. Perspectives, 32(5), 961-976. https://doi.org/10.1080/0907676X.2023.2219850

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv. https://doi.org/10.48550/arXiv.1609.08144

Zhang, Z. V., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. Assessing Writing, 36, 90-102. https://doi.org/10.1016/j.asw.2018.02.004