IREELL
Indonesian Review of English Education, Linguistics, and Literature
P-ISSN:3025-2369 I E-ISSN:3021-8101
https://jurnalfaktarbiyah.iainkediri.ac.id/index.php/ireel/index

Acceptability of English-Indonesian Machine Translation in Translating Stand-Up Comedy Humor

Reydina Isnata Churulliain*
Universitas Islam Negeri Sayyid Ali Rahmatullah Tulungagung
reydina08@gmail.com

Nany Soengkono Madayani
Universitas Islam Negeri Sayyid Ali Rahmatullah Tulungagung

**Corresponding author*

ABSTRACT

Despite the growing accuracy of machine translation (MT) in general-purpose tasks, its performance on culturally embedded, humor-driven text remains largely unexamined, particularly for the English-Indonesian language pair. This study examines the acceptability of humor translation produced by three MT tools, Google Translate, DeepL, and ChatGPT, using 20 humorous utterances drawn from Paul Snyder's stand-up comedy special "I Left the House for This" as the source data. A qualitative-dominant mixed-methods design was employed: the utterances were identified using Ruch, Attardo, and Raskin's General Theory of Verbal Humor, translated by the three tools, and evaluated using an adapted acceptability rubric, followed by a focus group discussion to consolidate the scores. On a 1-3 acceptability scale, Google Translate scored 1.42, DeepL 1.58 (both Unacceptable, though DeepL approached the Less Acceptable range) and ChatGPT 2.35, narrowly crossing into the Acceptable category. The main factors lowering acceptability were the use of formal written register in an informal spoken context, untransferred cultural references, and inconsistent pronoun choice. Despite occasional inconsistencies, ChatGPT produced the most natural and colloquial translations among the three tools. This study concludes that ChatGPT is currently the most suitable MT tool for translating humor from English stand-up comedy into Indonesian, while Google Translate and DeepL still require substantial human post-editing for this type of content. These findings offer novel empirical insight into MT-humor acceptability for an understudied language pair and genre.

Keywords: acceptability, humor translation, large language model (LLM), machine translation, stand-up comedy

Copyright: © 2026 by Authors.
Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Introduction

Translation is not merely a process of substituting words from one language into another; it is an act of transferring meaning and communicative intent across linguistic and cultural boundaries (Larson, 1984; Newmark, 1988). Contemporary translation studies increasingly view translation as a purposeful activity that must consider audience, function, and cultural context rather than linguistic correspondence alone (Nord, 1997). Within this functionalist perspective, the rapid growth of machine translation (MT) technology has reshaped how translation is practiced. MT refers to the automatic rendering of text between languages using computational systems (Gaspari & Hutchins, 2007), and its use has expanded considerably in educational and everyday contexts (Utimadini, 2023). Among the most widely used MT systems today are Google Translate, DeepL, and ChatGPT, each representing a different generation of translation technology, moving progressively from statistical and neural approaches toward generative language modeling. Google Translate initially relied on statistical machine translation before transitioning to a neural MT architecture, offering broad accessibility and multilingual coverage. DeepL is built on a neural MT architecture using transformer-based models, which allows it to capture deeper contextual relationships between words and produce more fluent and idiomatic output. ChatGPT, developed by OpenAI, represents a further shift toward large language models (LLMs): rather than mapping text through parallel corpora alone, it generates translations by drawing on broader contextual and generative language modeling, which allows for a more nuanced, context-sensitive understanding of meaning (Chan & Tang, 2024; Sahari et al., 2023).

Despite these advancements, MT systems continue to struggle with creative and culture-dependent text types such as humor, whose meaning and comic effect depend heavily on context, culture, and pragmatic inference. In translation quality assessment, acceptability has emerged as one of the most important indicators of translation success, referring to the extent to which a translation conforms to the linguistic norms, cultural expectations, and naturalness of the target language (Nababan et al., 2012). A translation is considered acceptable when it reads naturally and does not feel like a translation at all. This principle is particularly relevant to humor, since, as Low (2011) succinctly puts it, "if a joke is not translated

as a joke, the translation is bad" (p. 64). A translation is considered acceptable when it reads naturally and does not feel like a translation at all. Studies of Indonesian subtitles similarly show that translators tend to favor communicative translation, since it conveys a clear and natural message to the target audience (Arroihana & Kurniarahman, 2024). Even a semantically accurate translation fails as humor if the target audience cannot recognize its comedic intent, which means that in humor translation, acceptability matters more than literal accuracy or surface readability.

Humor itself is one of the most culturally embedded and linguistically complex phenomena to translate. Ruch, Attardo, and Raskin's General Theory of Verbal Humor (GTVH) explains that humor arises from an interplay of knowledge resources, most notably script opposition and logical mechanism, that depend on shared cultural background between speaker and audience (Ruch et al., 1993). Stand-up comedy exemplifies this complexity particularly well. Unlike written jokes, which exist as fixed, self-contained texts, stand-up comedy is inherently live and performative, relying on spontaneity, timing, and direct audience interaction to generate its comic effect (Double, 2014). This distinction makes it a more realistic representation of how humor functions in everyday spoken communication, compared to the scripted or written humor typically examined in earlier translation studies (Mintz, 1985). Translating such spontaneous, culturally loaded humor, whether by humans or machines, is therefore inherently demanding. Human translators typically rely on creative strategies such as adaptation and reformulation to preserve comic effect (Zabalbeascoa, 1996), whereas MT systems tend to process language more literally, often missing the pragmatic cues and cultural connotations that trigger laughter (Abu-Rayyash, 2024).

Previous research has begun to explore this issue from several angles. Ardi et al. (2022) found that MT-translated humorous texts could convey meaning but often lacked naturalness, requiring post-editing. Rudenko and Semenova (2024) similarly reported that ChatGPT-translated English-Russian puns tended to be literal and less humorous than human translations, while Calvo-Ferrer (2024) found that audiences often could not distinguish between human- and ChatGPT-translated subtitles, suggesting that AI has approached near-human fluency even if pragmatic sensitivity still lags behind. Chan and Tang's (2024)

systematic review of GPT-based translation likewise found that large language models frequently reach human-level quality for complex and creative texts, including humor, although output quality depends considerably on prompt design. However, very few of these studies address humor translation specifically through the lens of acceptability, and to the best of our knowledge, none has examined this question using authentic, spoken humor from stand-up comedy translated into Indonesian.

Beyond its theoretical relevance, this issue also carries practical significance for translation and English language education. As MT tools become more fluent and more widely used by students for academic and everyday purposes, it becomes increasingly important to examine not only whether they produce grammatically correct output, but whether that output would actually be accepted as natural by target-language audiences, particularly for text types, like humor, where naturalness is the very measure of success. This is especially relevant for translation students: understanding why certain machine-translated jokes are judged unnatural or culturally inappropriate can help them develop a deeper sensitivity to register, cultural context, and pragmatic meaning, competencies that are essential not only for professional translation but for intercultural communication more broadly. For English teachers, knowing which MT tool tends to produce more acceptable output can similarly inform decisions about incorporating humor and authentic spoken material, such as stand-up comedy, into classroom activities that build students' pragmatic and cross-cultural awareness.

This study addresses that gap by examining the acceptability of humor translation from English into Indonesian as produced by Google Translate, DeepL, and ChatGPT, using humorous utterances drawn from Paul Snyder's stand-up comedy special "I Left the House for This" as authentic data. Specifically, this research seeks to answer two questions: (1) how acceptable are the humor translations produced by the three MT tools, and (2) which tool performs best overall in preserving the intended meaning and comedic effect of the source text. By focusing on acceptability rather than accuracy alone, and by examining an authentic, spoken humor genre in the English-Indonesian language pair that has not previously been studied in this way, this study offers a novel empirical contribution to translation studies,

English language teaching, and the broader discussion of AI-mediated intercultural communication.

Methods

This study employed a qualitative-dominant mixed-methods design (Creswell, 2014; Johnson, Onwuegbuzie, & Turner, 2007), in which qualitative content analysis and interpretation formed the primary basis for addressing the research questions, while quantitative scoring and statistical comparison served a supporting role in identifying and substantiating patterns across the three MT tools. This design was considered appropriate for providing an in-depth interpretation of the linguistic and pragmatic features of translated humorous texts, while also allowing patterns in acceptability across tools to be systematically compared. The design allowed the researcher to explore how three MT tools, Google Translate, DeepL, and ChatGPT, rendered humor from English into Indonesian and to what extent their outputs were judged acceptable by native Indonesian speakers. Google Translate and DeepL were accessed via their free web-based versions. ChatGPT was accessed via its free tier on April 7, 2026, using the GPT-5.3 Instant model.

The data consisted of 20 humorous utterances taken from Paul Snyder's stand-up comedy special "I Left the House for This," published on 7 August 2024 through Dry Bar Comedy, a curated stand-up comedy platform on YouTube with more than three million subscribers. This source was selected both theoretically and empirically: theoretically, Snyder's observational and relational style of humor, built on informal register and everyday pragmatic inference rather than wordplay, provides an ideal test case for evaluating naturalness in translation; empirically, the video's high engagement (approximately 2.2 million views and audience comments reflecting broad cross-cultural resonance) confirms its effectiveness as an authentic humor sample (Dry Bar Comedy, 2024). A purposive sampling technique was applied to select the data: the full video transcript was examined, and utterances were identified as humorous units if they exhibited elements central to the GTVH framework (Ruch, Attardo, & Raskin, 1993), namely incongruity or script opposition, wordplay, or pragmatic irony. This process yielded 20 humorous utterances, which were then translated into

Indonesian by the three MT tools, producing three parallel sets of target texts for comparison. As the video is publicly available on YouTube, its transcript was used in this study solely for non-commercial academic research purposes, consistent with fair-use principles for scholarly analysis of publicly accessible media.

In this study, the researcher served as the primary instrument, responsible for identifying the humorous utterances and generating the MT outputs. The researcher did not act as a rater; acceptability judgments were made independently by three separate raters, described below, to preserve the objectivity of the evaluation. As a supporting instrument, the study adapted the Translation Quality Assessment rubric originally developed by Nababan et al. (2012), which in its original form evaluates translation quality across three weighted dimensions: accuracy, acceptability, and readability. This study adopted only the acceptability dimension, as this dimension aligns most directly with the study's focus on the naturalness and cultural appropriateness of humor translation; the accuracy and readability dimensions were excluded, as they fall outside the scope of this study. The three-point scale used to assess acceptability was adopted exactly as published in the original instrument, without modification: 3 (Acceptable) for a natural translation that follows Indonesian linguistic and cultural norms and does not feel like a translation; 2 (Less Acceptable) for a generally understandable translation containing minor unnatural expressions; and 1 (Unacceptable) for a translation that sounds awkward or inconsistent with Indonesian norms. Because acceptability was scored by three raters per item, item-level results were reported as mean scores rather than single categorical ratings; these means were subsequently converted back into the three original categories using cutoffs derived fromequal division of the 1–3 scale range: Acceptable (≥ 2.34), Less Acceptable (1.67–2.33), and Unacceptable (< 1.67).

Data were collected through content analysis and a focus group discussion (FGD). Content analysis was conducted by reading through the full transcript and identifying humorous units based on two coding criteria drawn from the GTVH (Ruch, Attardo, & Raskin, 1993): the presence of a clear script opposition, a clash between two incompatible interpretations, and an accompanying logical mechanism that resolves or heightens that clash. Utterances meeting both criteria were coded according to their specific script

opposition and logical mechanism before being translated by each MT tool; the resulting 20 humorous units averaged 25.9 words (130.9 characters) in length, ranging from 9 to 60 words. Subsequently, three raters, selected for their background in translation studies, proficiency in both English and Indonesian, and understanding of translation quality assessment, were invited to participate. Verbal informed consent was obtained from all three raters prior to the FGD, which was conducted in English at the raters' preference. Prior to reaching consensus, the raters independently scored all 60 translated outputs (20 utterances × 3 MT tools); inter-rater agreement at this stage was slight (Fleiss' $\kappa = .161$), with all three raters assigning an identical score in 21.7% of cases. This level of agreement is consistent with the inherently subjective nature of naturalness judgments and reflects the value of the subsequent FGD, in which the 23-minute discussion allowed raters to reconcile differing interpretations and arrive at a consolidated score for each item.

Data analysis followed Miles et al. (2014) three-stage model of data condensation, data display, and conclusion drawing: humorous utterances and their three MT translations were condensed and organized into comparative tables, after which patterns in acceptability scores and their contributing linguistic factors were interpreted to determine which MT tool produced the most acceptable output overall. To establish trustworthiness, this study addressed all four criteria proposed by Lincoln and Guba (1985). Credibility was established through triangulation of MT outputs and multiple raters, member checking with the raters, and peer debriefing with the research supervisor. Dependability and confirmability were supported by maintaining a complete audit trail, including audio recordings and verbatim transcripts of the FGD session and individual rater scoring sheets, allowing the analytic process and findings to be traced back to the raw data. Transferability was addressed through thick description of the data source and selection criteria, including detailed contextual information about the stand-up comedy performance and the GTVH-based criteria used to identify humorous utterances, enabling readers to judge the applicability of the findings to similar contexts.

Result and Discussion

In response to RQ1, concerning how acceptable the humor translations produced by the three MT tools are, the evaluation of the 20 humorous utterances shows clear differences among them. Google Translate obtained the lowest overall mean score of 1.42, placing it in the Unacceptable category; DeepL scored slightly higher at 1.58, also categorized as Unacceptable; while ChatGPT achieved a mean score of 2.35, the only tool to reach the Acceptable threshold. It should be noted that this score exceeds the threshold by a narrow margin (2.35 against a cutoff of 2.34), and this categorical distinction should therefore be interpreted with appropriate caution. Turning to RQ2, on which tool performs best overall, ChatGPT scored highest on 17 of the 20 items and produced the single highest-scoring datum in the dataset (a joke about sharing clothes with a sibling, mean score 3.00, unanimously rated Acceptable by all three raters), whereas Google Translate and DeepL each had a majority of items falling into the Unacceptable range. A Friedman test confirmed that this difference across the three tools was statistically significant, χ^2 (2) = 25.44, $p < .001$. Post-hoc Wilcoxon signed-rank tests with Bonferroni correction showed that ChatGPT scored significantly higher than both Google Translate ($p < .001$) and DeepL ($p < .001$), while the difference between Google Translate and DeepL was not statistically significant ($p = .193$), indicating that these two tools performed comparably poorly relative to ChatGPT. Table 1 summarizes the overall comparison across all evaluative criteria.

Table 1

Overall Comparison of the Three Machine Translation Tools

Criteria	Google Translate	DeepL	ChatGPT
Overall mean score	1.42	1.58	2.35
Overall category	Unacceptable	Unacceptable	Acceptable
Items rated Acceptable (mean ≥ 2.34)	0 (0%)	0 (0%)	4 (20%)
Items rated Less Acceptable (1.67–2.33)	7 (35%)	9 (45%)	15 (75%)
Items rated Unacceptable (< 1.67)	13 (65%)	11 (55%)	1 (5%)
Highest-scoring item	mean 2.33	mean 2.33	mean 3.00

To illustrate how these scores were derived, Table 2 presents three representative items alongside their source text, the three MT outputs, and the raters' scores. Datum 1 is the comedian's opening joke asking for applause simply for leaving the house as a self-proclaimed introvert; Datum 14 is the item that received the highest overall score in the dataset, a joke about clothes-sharing with a sibling; and Datum 20, the closing joke of the routine referencing Google Maps search suggestions, was the item on which ChatGPT scored lowest due to an added, unwarranted geographic detail.

Table 2

Sample Humorous Utterances and Their Translations by the Three MT Tools

Datum	MT tool	Text	Score
Datum 1	Source text	"*...also give me a round of applause because I am a huge introvert and I left the house for this today.*"	
	Google Translate	"Sekarang, jika Anda tidak keberatan, beri saya tepuk tangan juga karena saya seorang introvert sejati dan saya keluar rumah untuk ini hari ini."	M = 1.00 (Unacceptable)
	DeepL	"Nah, kalau boleh, tolong beri saya tepuk tangan juga karena saya orang yang sangat introvert dan hari ini saya berani keluar rumah untuk acara ini."	M = 1.67 (Less Acceptable)
	ChatGPT	"Kalau kalian nggak keberatan, tolong kasih tepuk tangan juga buat saya—karena saya ini introvert parah dan hari ini saya berhasil keluar rumah."	M = 2.33 (Less Acceptable)
Datum 14	Source text	"*We share clothes It is just a lot more embarrassing for her.*"	
	Google Translate	"Aku dan anakku yang masih SMA, kami memiliki ukuran tubuh yang sama, jadi kami bisa berbagi pakaian, yang menurutku sangat keren karena menghemat uang. Hanya saja itu jauh lebih memalukan baginya."	M = 1.67 (Less Acceptable)
	DeepL	"Saya dan anak SMA saya, kami memiliki ukuran yang sama, jadi kami bisa berbagi pakaian, yang menurut saya sangat keren karena menghemat uang. Hanya saja itu jauh lebih memalukan baginya."	M = 1.00 (Unacceptable)
	ChatGPT	"Kami suka tukar-tukeran baju... cuma buat dia jauh lebih memalukan."	M = 3.00 (Acceptable)

Datum 20	Source text	"*Google Maps showed me five places where I could dispose of a body.*"	
	Google Translate	"Google Maps menunjukkan kepada saya lima tempat di mana saya bisa membuang mayat."	M = 2.00 (Less Acceptable)
	DeepL	"Google Maps menunjukkan lima tempat di mana saya bisa membuang mayat."	M = 2.33 (Less Acceptable)
	ChatGPT	"Google Maps bahkan nunjukin lima lokasi buat 'buang mayat' antara Salt Lake City ke Provo."	M = 1.33 (Unacceptable)

Note. Adapted from the acceptability rubric of Nababan, Nuraeni, and Sumardiono (2012); scores are FGD mean ratings from three raters (1 = Unacceptable, 2 = Less Acceptable, 3 = Acceptable).

These results indicate that all three tools struggled to consistently produce acceptable humor translations, but the gap between conventional neural MT (Google Translate, DeepL) and the generative large language model (ChatGPT) is substantial. This finding echoes House's (2014) argument that translation quality must ultimately be judged by whether the target text fulfills its communicative purpose, which, in this case, is to make Indonesian listeners laugh. When translations came out stiff and overly formal, they failed to serve this purpose regardless of whether the underlying meaning was technically correct, aligning with Low's (2011) observation that a translated joke which no longer functions as a joke cannot be considered a good translation.

One recurring problem across all three tools was the use of formal written register in a context that demanded informal, spontaneous-sounding speech. For instance, in the comedian's opening line asking for applause as a self-proclaimed introvert, Google Translate produced a stiff construction using the formal conditional "jika Anda tidak keberatan" and the pronoun "Anda," earning a unanimous Unacceptable score. As one rater explained during the FGD, stand-up comedy in Indonesian rarely uses the formal pronoun saya, relying instead on informal forms such as gue or aku. ChatGPT's version of the same line used markedly more colloquial phrasing, such as "introvert parah" (roughly, "a hopeless introvert"), and was rated more favorably, though even ChatGPT retained the formal pronoun saya inconsistently, an issue discussed further below. This pattern supports Mintz's (1985) characterization of stand-up comedy as a form of direct, unscripted-sounding address between comedian and

audience: when this spoken register is rendered into formally structured written Indonesian, the sense of intimacy that drives the humor is lost, a challenge also noted by Díaz Cintas and Remael (2014) in their discussion of audiovisual humor and comedic timing.

A second major factor lowering acceptability was the failure to adapt culturally specific references. Data involving culturally bound content, such as a comparison between introverts during the COVID-19 pandemic and soldiers storming the beaches of Normandy, or a reference to Werther's Original candies, received consistently low scores because none of the three tools attempted cultural localization; they simply retained the source-language reference. According to the GTVH, humor depends on script opposition that relies on shared cultural knowledge between speaker and audience (Ruch et al., 1993); when that shared knowledge is absent, as is the case for many Western pop-cultural references among Indonesian audiences, the humor fails to register regardless of translation accuracy. This tendency to preserve the source reference rather than adapt it mirrors Venuti's (2003) concept of foreignization, which, while sometimes appropriate in other translation contexts, tends to work against humor, where audience accessibility is essential for a joke to land. Chiaro (2010) similarly frames humor translation as an act of negotiation between two linguistic communities, one that machine translation, in its current form, is not yet equipped to perform.

Analyzing the data through the six knowledge resources of the GTVH (Ruch, Attardo, & Raskin, 1993) clarifies why certain translations preserved humor while others did not. Each utterance in this study was purposively selected because it carried a clear script opposition, the clash between two incompatible interpretations, paired with a specific logical mechanism that resolves or heightens that clash. In Datum 1, for instance, the script opposition lies between introverted behavior and social expectation, resolved through exaggeration and contrast; Google Translate's overly formal rendering flattened this contrast by removing the casual self-deprecation central to the joke, whereas ChatGPT's more colloquial phrasing preserved it, earning a higher score. A similar pattern appears in Datum 10, where the script opposition between marriage and criminal captivity relies on metaphorical exaggeration; reconstructing this joke required recognizing the implicit "blink twice if you're in danger" signal embedded in the source text and finding a way to convey it in Indonesian, a challenge

discussed further below in comparison across the three tools. These examples suggest that acceptable humor translation depends not merely on lexical accuracy but on whether the target text reconstructs the specific knowledge resources, particularly script opposition and logical mechanism, that generate the comic effect in the source text.

ChatGPT's comparatively strong performance may be attributed in part to its generative architecture, which, unlike conventional neural MT systems that primarily map source text to target text through parallel corpora, is designed to produce more expressive and contextually adapted output (Chan & Tang, 2024). Its translations frequently used natural colloquial expressions, such as kena banget ("that really hits") and nggak bakal ngobrolin ("won't talk about it"), reflecting a level of sensitivity to informal Indonesian register that neither Google Translate nor DeepL consistently achieved. In one item, ChatGPT even performed a limited form of cultural adaptation, translating the English convention "blink twice if you're in danger" into the Indonesian equivalent kode SOS, making the reference accessible to audiences unfamiliar with the source convention. By contrast, Google Translate rendered the line as "Aku nggak bisa berkedip lebih cepat lagi" and DeepL as "Saya tidak bisa berkedip lebih cepat lagi," both translating "blinking" literally without conveying the implicit danger signal at all, leaving the joke's underlying logic inaccessible to readers unfamiliar with the English convention. ChatGPT's adaptation is thus a rare instance of the kind of creative cultural mediation Chiaro (2010) associates with successful humor translation. These findings align with Chan and Tang (2024) and Abu-Rayyash (2024), both of whom found that GPT-based translation can approach or match human-level quality for creative and culturally nuanced texts, and with Chen and Lin (2025), who similarly found ChatGPT to outperform Google Translate and DeepL in cultural sensitivity when translating Chinese tourism texts.

Even so, ChatGPT was not without weaknesses. Its most consistent limitation was the tension between informal vocabulary and its persistent use of the formal first-person pronoun saya, even in otherwise highly colloquial sentences, an inconsistency one rater described as making the translation "feel like you want to joke around, but you used saya, so it's weird." ChatGPT also occasionally added information not present in the source text, such as an unwarranted geographic detail in one item, echoing Sun's (2024) observation that ChatGPT's

creative translation style can sometimes come at the cost of fidelity. DeepL, meanwhile, showed modest improvements over Google Translate in vocabulary choice, using more natural equivalents such as ngobrol instead of the stiffer membicarakannya, but it still defaulted to formal sentence structures in most items and, in one case, weakened a joke by translating "stormed" as mendarat ("landed"), removing the aggressive imagery central to the joke's absurdity. Google Translate, for its part, performed adequately only on short, factually simple utterances with little cultural or humorous loading, such as a straightforward line about celebrating a wedding anniversary, but consistently failed on data requiring colloquial phrasing or cultural inference.

Taken together, these findings confirm and extend earlier research on MT and humor translation. Whereas Ardi et al. (2022) found that conventional MT tools could convey the surface meaning of humorous texts but lacked naturalness, and Rudenko and Semenova (2024) reported that even ChatGPT tended toward overly literal, formally registered output when translating English-Russian puns, the present study suggests that ChatGPT has made meaningful progress in producing naturalistic Indonesian output, though this progress is uneven at the level of individual utterances. The results also point to a gap in the existing MT literature, much of which appears to focus on European language pairs: informal spoken Indonesian, with its rich discourse markers, sharp formal-colloquial register distinctions, and complex pronoun system, poses translation challenges that are rarely addressed in current MT research and training data. This suggests that evaluating MT systems for register sensitivity and cultural appropriateness in specific, understudied language pairs, rather than general translation accuracy alone, is an important direction for future work.

These findings also carry implications beyond translation theory. Acceptability, as measured in this study, reflects whether a translation would be recognized by its target audience as natural rather than machine-generated; when even the best-performing tool falls short of this standard, its output cannot yet be considered ready for use without further human revision. Presenting unedited MT translations as finished work in such cases, particularly in academic or professional settings, therefore raises concerns about intellectual honesty and quality control. This reinforces House's (2014) point that translation quality

assessment must consider not only linguistic form but also the responsibility of whoever delivers the final text to its intended audience. For educators, this suggests that MT tools are best introduced not as replacements for human judgment but as drafting aids whose output must be critically reviewed, an approach that is especially important when the source material, like stand-up comedy, depends so heavily on register, timing, and cultural inference for its effect.

Conclusion

This study set out to examine how acceptable Google Translate, DeepL, and ChatGPT are in translating humor from English into Indonesian, and which of the three performs best overall. The findings show that Google Translate and DeepL both fell into the Unacceptable category, primarily due to their reliance on formal written register, failure to adapt culturally specific references, and inconsistent pronoun use, while ChatGPT was the only tool to reach the Acceptable category, owing to its more expressive and contextually sensitive output. Even so, ChatGPT's performance was not uniform across all items, and its remaining weaknesses in pronoun consistency and fidelity indicate that no MT tool tested in this study can yet be relied upon to produce humor translations that are acceptable without further human revision.

Beyond confirming that MT tools continue to struggle with creative, culturally loaded text types, this study extends the concept of acceptability (Nababan et al., 2012) to a previously underexamined genre, authentic spoken stand-up comedy, and to a previously underexamined language pair, English-Indonesian. In doing so, it provides preliminary support for the broader claim that large language models represent a qualitative advance over conventional neural MT systems in handling pragmatically complex content, while also highlighting the persistent limits of automated translation in matters of register and cultural nuance.

This study is limited by its relatively small dataset of 20 items drawn from a single comedian and by its focus on acceptability alone, excluding the related dimensions of accuracy and readability. Future research may address these limitations by expanding the dataset across multiple comedians and comedic styles, incorporating a human translator's

version as a benchmark, and examining whether targeted prompt engineering, for instance, explicitly instructing ChatGPT to use informal pronouns and avoid formal discourse markers, can further improve the acceptability of AI-generated humor translation. For practical purposes, the findings suggest that students and language practitioners should treat MT output, including from ChatGPT, as a first draft requiring critical human revision rather than a finished product, particularly when translating informal, humor-dependent, or culturally embedded content.

References

Abu-Rayyash, H. (2024). AI meets comedy: Viewers' reactions to GPT-4 generated humor translation. *Ampersand*, *12*, 100162. https://doi.org/10.1016/j.amper.2023.100162

Ardi, H., Al Hafizh, M., Rezqi, I., & Tuzzikriah, R. (2022). Can machine translations translate humorous texts? *Humanus*, *21*(1), 99–112. https://doi.org/10.24036/humanus.v21i1.115698

Arroihana, F. Z., & Kurniarahman, I. (2024). Analyzing Translation Methods in the Hidden Figures Movie. *IREELL, Indonesian Review of English Education, Linguistics, and Literature, 2*(2), 137–160. https://doi.org/10.30762/ireell.v2i2.3720

Calvo-Ferrer, J. R. (2024). Can you tell the difference? A study of human vs machine-translated subtitles. *Perspectives*, *32*(6), 1115-1132. https://doi.org/10.1080/0907676X.2023.2268149

Chan, V., & Tang, W. K. W. (2024). GPT for translation: A systematic literature review. *SN Computer Science*, *5*(8), 986. https://doi.org/10.1007/s42979-024-03340-z

Chen, S., & Lin, Y. (2025). A multidimensional comparison of ChatGPT, Google Translate, and DeepL in Chinese tourism texts translation: Fidelity, fluency, cultural sensitivity, and persuasiveness. *Frontiers in Artificial Intelligence*, *8*, 1619489. https://doi.org/10.3389/frai.2025.1619489

Chiaro, D. (2010). *Translation, humour and literature* (Vol. 1). Continuum.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.

Díaz Cintas, J., & Remael, A. (2014). *Audiovisual translation: Subtitling*. Routledge.

Double, O. (2014). *Getting the joke: The inner workings of stand-up comedy* (2nd ed.). Bloomsbury Methuen Drama.

Dry Bar Comedy. (2024). *Paul Snyder: I Left the House for This*. YouTube. https://youtu.be/8DNQ8DYgCJE

Gaspari, F., & Hutchins, W. J. (2007). Online and free! Ten years of online machine translation: Origins, developments, current use and future prospects. *Proceedings of MT Summit XI*, 199–206.

House, J. (2014). *Translation quality assessment: Past and present*. Routledge.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, *1*(2), 112–133. https://doi.org/10.1177/1558689806298224

Larson, M. L. (1984). *Meaning-based translation: A guide to cross-language equivalence*. University Press of America.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage Publications.

Low, P. A. (2011). Translating jokes and puns. *Perspectives: Studies in Translatology*, *19*(1), 59-70. https://doi.org/10.1080/0907676X.2010.493219

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.

Mintz, L. E. (1985). Standup comedy as social and cultural mediation. *American Quarterly*, *37*(1), 71. https://doi.org/10.2307/2712763

Nababan, M., Nuraeni, A., & Sumardiono. (2012). Pengembangan model penilaian kualitas terjemahan. *Kajian Linguistik dan Sastra, 24*(1), 39–57.

Newmark, P. (1988). *A textbook of translation*. Prentice Hall.

Nord, C. (1997). *Text analysis in translation: Theory, methodology, and didactic application of a model for translation-oriented text analysis*. Rodopi.

Ruch, W., Attardo, S., & Raskin, V. (1993). Toward an empirical verification of the General Theory of Verbal Humor. *Humor*, *6*(2), 123–136. https://doi.org/10.1515/humr.1993.6.2.123

Rudenko, E. S., & Semenova, M. Y. (2024). Artificial vs human intelligence: a case study of translating jokes based on wordplay. *Research Result. Theoretical and Applied Linguistics*, *10*(4), 53–93. https://doi.org/10.18413/2313-8912-2024-10-4-0-4

Sahari, Y., Al-Kadi, A. M. T., & Ali, J. K. M. (2023). A cross-sectional study of ChatGPT in translation: Magnitude of use, attitudes, and uncertainties. *Journal of Psycholinguistic Research*, *52*(6), 2937–2954. https://doi.org/10.1007/s10936-023-10031-y

Sun, R. (2024). Evaluating the translation accuracy of ChatGPT and DeepL through the lens of implied subjects. *AWEJ for Translation & Literary Studies*, *8*(4).

Utimadini, N. J. (2023). Exploring perceptions of machine translation as a tool for EFL learning.

Jurnal Pendidikan Bahasa, *12*(2), 484–494. https://doi.org/10.31571/bahasa.v12i2.7128

Venuti, L. (2003). *The translator's invisibility*. Routledge. https://doi.org/10.4324/9780203360064

Zabalbeascoa, P. (1996). Translating jokes for dubbed television situation comedies. *The Translator*, *2*(2), 235–257. https://doi.org/10.1080/13556509.1996.10798976