Compilation and Translation Review
Vol. 19, No. 2 (September 2026), 189-252
DOI: 10.29912/CTR.202609_19(2).0007

Untranslatables in Taiwanese: A Corpus-Based Study of Semantic Leakage in AI Translations

Vincent Chieh-Ying Chang

This study proposes a multidimensional metric—Semantic Leakage Indicator (SLI) —to quantify meaning loss in AI-generated translations of Taiwanese into Mandarin. Drawing on a corpus of 26,046 Taiwanese sentences sourced from diverse spoken and written contexts, the analysis identifies five morpho-pragmatic features highly susceptible to semantic erosion: aspect markers, sentence-final particles, reduplication, kinship terms, and directional compounds. The SLI integrates entropy reduction, back-translation divergence, and semantic embedding shifts to detect untranslatable constructions that evade surface-level fluency metrics. Results reveal that even when lexical equivalents are present, high-SLI instances exhibit pragmatic distortion and affective dissonance. Cluster analysis delineates patterns of leakage across grammatical categories, while qualitative examples illustrate genre-specific vulnerabilities. The study advances translation studies by offering a scalable, language-sensitive diagnostic tool for evaluating AI performance on low-resource, morphologically rich languages. It contributes empirically to discussions on untranslatability, pragmatics, and meaning preservation in AI translation, with direct implications for algorithm design, language preservation, and corpus-informed evaluation in underrepresented languages.

Keywords: untranslatability, semantic leakage, Taiwanese, low-resource language, AI translation

Received: June 27, 2025
Revised: October 28, 2025; January 13, 2026; January 25, 2026; March 26, 2026
Accepted: June 3, 2026

Vincent Chieh-Ying Chang, Associate Professor, Department of English, Tamkang University, E-mail: 159738@o365.tku.edu.tw

This article is dedicated to the memory of my late father, whose everyday use of Taiwanese in our home first revealed to me how language carries memory, emotion, and identity. His voice continues to inform my thinking about what translation as accessibility is, and this study is, in many ways, an extended act of listening back to his stories. Simultaneously, I extend profound and heartfelt gratitude to the anonymous reviewers, whose discerning insights and constructive counsel have refined and fortified the manuscript.

編譯論叢
第十九卷　第二期（2026年9月），189—252

臺語中之不可譯項：AI 翻譯中語意流失的語料庫研究

張介英

本研究提出一項多維度指標——語意流失指數（Semantic Leakage Indicator, SLI），用以量化臺語翻譯為華語時，人工智慧譯文中之語意流失。研究使用之語料庫共計26,046組臺語句子，涵蓋多元的口語與書面語境。分析聚焦於五項高度易受語意侵蝕之形態語用特徵：體標記、句尾助詞、重疊形式、親屬稱謂詞，以及動向複合詞。SLI 結合語意熵減、回譯差異及語意嵌入位移三項演算法，以識別那些表層流暢度指標難以偵測之「不可譯」結構。研究結果顯示，即便譯文表面上存在詞彙對應，高 SLI 值案例仍常伴隨語用扭曲與情感不協調。群聚分析則勾勒出不同語法範疇中的語意流失樣態；另輔以質性實例，呈現各文體類型所固有之翻譯脆弱性。此研究為翻譯學提供一套可擴充、具語言敏感度之診斷工具，用以評估人工智慧在低資源、形態複雜語言上的表現，並於「不可譯性」、語用意涵與語意保存等議題上，貢獻具體實證價值，對演算法設計、語言保存與語料本位評估等均具實質啟示意義。

關鍵詞：不可譯性、語意流失、臺語、低資源語言、AI 翻譯

收件：2025 年 6 月 27 日
修改：2025 年 10 月 28 日、2026 年 1 月 13 日、2026 年 1 月 25 日、2026 年 3 月 26 日
接受：2026 年 6 月 3 日

張介英，淡江大學英文學系副教授，E-mail: 159738@o365.tku.edu.tw。

Introduction

Translation scholars have long contested the assumption that "everything is translatable," a challenge that becomes particularly acute when AI systems confront low-resource languages such as Taiwanese (Tâi-gí) (臺語), which lacks standardized orthography and robust digital resources and is therefore structurally and culturally underrepresented in training data; even amid major gains in neural machine translation, the theoretical and practical problem of what remains "untranslatable" persists. Building on this concern, this study operationalizes semantic leakage—the distortion, omission, generalization, or pragmatic/stylistic reconfiguration of meaning in AI-mediated transfer—and proposes a corpus-based method that moves beyond reference- or human-dependent evaluation (e.g., BLEU, METEOR) by using back-translation divergence to systematically surface patterns of untranslatability. Using a corpus of 26,046 Taiwanese sentences translated to Mandarin and back-translated to Taiwanese with ChatGPT-4o, we quantify meaning loss and identify linguistic features that reliably resist faithful transfer, integrating these signals into a quantitative metric, the Semantic Leakage Indicator (SLI),[1] complemented by qualitative analysis. The contribution is both disciplinary and practical: It links longstanding theorization—from Catford's (1965) linguistic versus cultural untranslatability, through spectrum-based reconceptualizations and Venuti's (2017) foreignization/domestication, to Apter's (2019) political framing of untranslatability as resistance—to the empirical realities of contemporary AI translation, clarifying where and how current models systematically fail under conditions of linguistic diversity.

[1] The Semantic Leakage Indicator (SLI) is a normalized composite score that quantifies semantic, pragmatic, and stylistic divergence across a three-step translation chain (source → translation → back-translation); higher values indicate greater leakage arising from cross-linguistic untranslatability and/or model-specific limitations.

My research questions are threefold: (a) What specific patterns of untranslatability emerge when AI systems translate between Taiwanese and Mandarin Chinese, considering their linguistic and cultural proximity yet distinct features? (b) How can semantic leakage be effectively quantified and categorized using a combination of computational and qualitative methods, moving beyond simple error counts? (c) What linguistic features (lexical, syntactic, pragmatic, cultural) correlate most strongly with higher degrees of semantic leakage, and what does this reveal about the challenges faced by AI in handling low-resource languages? By addressing these questions, I contribute to the ongoing dialogue between translation theory and computational linguistics, while offering practical insights for improving AI translation systems for low-resource languages and fostering more equitable digital communication.

Literature Review

Theoretical Frameworks of Untranslatability: From Dichotomy to Spectrum

The concept of untranslatability has long haunted translation theory. Catford (1965) defined it in structural-linguistic terms, distinguishing between linguistic untranslatability "due to formal non-correspondence between language systems" (p. 94) and cultural untranslatability "arising from the absence of relevant situational features or concepts in the target culture" (p. 99). This binary framework, while foundational, has been significantly expanded by subsequent scholars. Nida (1964) introduced the influential distinction between formal equivalence "focusing on the message itself, in both form and content" (p. 159) and dynamic equivalence "aiming for equivalent effect on the target audience," (p. 159)

implicitly acknowledging that perfect translation might require prioritizing one type of equivalence over another. That said, Catford's (1965) concept of cultural untranslatability encompasses a broad spectrum of cultural phenomena, including idioms, proverbs, and culturally embedded linguistic practices, rather than merely isolated cultural terms. In my study, I operationalize a narrower category of "cultural-specific terms" (referring specifically to items related to Taiwanese customs, rituals, cuisine, and social practices) to enable more granular analysis of different types of untranslatability. Pym (2023) shifted the conversation toward the ethics of non-translation and translator agency, emphasizing situations where the translator chooses not to translate, or only partially does so, as a political or epistemological gesture. He argues that untranslatability is not merely a technical limitation but often a deliberate strategy that preserves the integrity of the source text (ST). Venuti (2017) framed untranslatability within the paradigm of "resistance," arguing that foreignizing strategies must preserve the strangeness of the original rather than domesticate its force. This perspective positions untranslatability as a positive value rather than a deficiency.

The descriptive translation studies (DTS) paradigm, as articulated by Toury (2012) and Even-Zohar (2005), moved away from prescriptive notions of translatability to examine how translations function within target cultures, regardless of their fidelity to STs. This approach acknowledges that what counts as "acceptable translation" varies across cultures and historical periods, further complicating the notion of untranslatability. Pym (2023) later observed that untranslatability might be better understood as a spectrum of resistance rather than a binary condition, with some elements requiring greater effort or creativity to translate than others (p. 136). More recently, Apter (2019) situated untranslatability within the geopolitics of reading and emphasized its political implications while proposing that some terms cannot, and perhaps should not, be

carried across linguistic boundaries, particularly in postcolonial or sacred contexts. This perspective aligns with what Venuti (2017) termed "foreignization," a strategy that deliberately preserves the otherness of the ST. Lai (2024) extends this argument specifically to Taiwanese contexts, noting that certain cultural-political terms resist translation precisely because their untranslatability serves as a form of resistance against linguistic hegemony.

Computational Approaches to Translation Quality and Divergence

Recent evaluations of GPT-4 on low-resource languages have shown similar performance patterns, with capability gaps widening as language resources decrease (Liang et al., 2025; Wang & Wang, 2025). Traditional machine translation evaluation metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE (Lin, 2004) have been criticized for their limited ability to capture semantic and pragmatic dimensions of translation quality (Callison-Burch et al., 2006; Reiter, 2018). These reference-based metrics primarily measure surface-level similarity between machine translations and human references, often failing to account for valid translation variations or deeper semantic equivalence. Recent work by Zhang et al. (2019) and Zhao et al. (2019) has attempted to address these limitations by incorporating contextual embeddings and neural metrics, but challenges remain in evaluating translations of low-resource languages where high-quality reference translations are scarce.

Round-trip translation (RTT),[2] which involves translating from a ST to a target text (TT) and back to the source language (BT), offers an alternative approach to

[2] Round-Trip Translation (RTT) refers to a translation procedure in which a source text is first translated into a target language and then translated back into the source language. RTT is employed here as a diagnostic tool to surface semantic leakage by comparing the original source sentence with its back-translated counterpart, thereby revealing shifts in meaning, pragmatics, or stylistic features introduced during the translation process.

assessing translation quality without requiring reference translations (Aiken & Park, 2010; Somers, 2005). While RTT has been criticized for potentially masking errors that persist across both translation directions, recent studies have demonstrated its utility in identifying specific types of translation problems, particularly when combined with embedding-based similarity measures (Peng, 2021). The back-translation approach has gained renewed attention in neural machine translation research, both as an evaluation technique and as a data augmentation strategy (Sennrich et al., 2015).

Computational stylometry, which applies statistical methods to analyze linguistic style, provides another lens for examining translation quality. Stylometric approaches have been used to detect translationese—the distinctive features that mark a text as a translation rather than an original (Volansky et al., 2015). Measures such as lexical diversity, syntactic complexity, and information density can reveal subtle shifts that occur during translation, even when the semantic content is preserved. These methods are particularly valuable for identifying cases where translations maintain denotational meaning but lose stylistic or pragmatic dimensions.

Taiwanese Language: Linguistic Features and Translation Challenges

Taiwanese (Tâi-gí) presents several distinctive linguistic features that pose challenges for machine translation systems. As a Sinitic language with significant historical influence from Japanese, Dutch, and indigenous Formosan languages, Taiwanese exhibits phonological, lexical, and grammatical patterns that diverge from Mandarin Chinese in important ways. The language features a complex tonal system (typically seven to eight tones compared to Mandarin's four, plus complex tone sandhi rules) (Hsieh, 2014; Zhang, 2014) and lacks a universally adopted standardized writing system, though Tâi-lô romanization (臺羅) has gained

prominence in academic contexts (Cheng, 2016; Hsieh, 2014). Additionally, Taiwanese presents a classic case of a low-resource language in computational linguistics. Despite being spoken by approximately 70% of Taiwan's population (Chang, 2022), Taiwanese faces significant resource limitations that affect Natural Language Processing (NLP) performance. In terms of limited parallel corpora, the largest publicly available Taiwanese-Mandarin parallel corpus contains approximately 27,000 sentence pairs (Liao, 2022), compared to millions available for high-resource languages. This corpus exhibits limited domain coverage, focusing primarily on educational materials and folk literature. As for orthographic inconsistency, Taiwanese lacks a universally adopted standardized writing system, with texts appearing in various romanization systems (Pe̍h-ōe-jī 白話字 , Tâi-lô 臺羅) and *Han* 漢 character variants, complicating corpus development and consistency (Liu et al., 2016). Concerning quality issues, existing parallel data often exhibits quality concerns, including non-native expressions, inconsistent dialectal usage, and alignment errors (Klöter, 2005). Regarding domain gaps, available corpora overrepresent formal registers and traditional contexts while underrepresenting contemporary usage, technical domains, and colloquial speech (Liu et al., 2016). These resource limitations directly impact the performance of AI translation systems for Taiwanese, potentially exacerbating semantic leakage beyond what would be observed with more robust resources.

Several specific linguistic features of Taiwanese create challenges for AI translation systems:

1. Modal and pragmatic particles: Taiwanese employs over 30 sentence-final particles that encode speaker stance, evidentiality, and interpersonal dynamics (e.g., "-leh," "-ho^{n}h," "-lah"). These particles often lack direct equivalents in Mandarin Chinese, leading to what Capone (2023) terms "pragmatic compression" (p. 1) in translation. The loss of these particles can significantly

alter the interpersonal force and pragmatic meaning of utterances, even when propositional content is preserved.

2. Reduplication patterns: Taiwanese uses various reduplication patterns to express aspectual meanings, intensity, distribution, and affect (e.g., "kóng-kóng-kiò-kiò," "liâm-liâm-á"). As Zeitoun and Wu (2006) demonstrates, these patterns resist straightforward translation into languages with different morphological systems. The semantic and pragmatic nuances conveyed through reduplication are often flattened or omitted entirely in translation.
3. Kinship terminology: Taiwanese maintains a more elaborate kinship terminology system than Mandarin, distinguishing maternal from paternal relatives and encoding age relationships among siblings and cousins (e.g., "a-kong" vs. "a-má," "a-hia^{n}" vs. "a-tī"). These distinctions often collapse in translation, resulting in what Sandel (2002) calls "kinship compression" (p. 257). The social and cultural information embedded in these terms is frequently lost in translation.
4. Aspect marking: Taiwanese employs a complex system of aspect markers that interact with verbal semantics in ways that differ significantly from Mandarin (e.g., "-leh" for progressive, "-kòe" for experiential). These differences create what Sharma and Deo (2009) describes as "aspectual misalignment" (p. 16) in translation. The temporal and aspectual precision of Taiwanese expressions is often compromised when translated into languages with different aspectual systems.
5. Sound symbolism and onomatopoeia: Taiwanese makes extensive use of sound symbolic expressions and onomatopoeia that carry semantic and affective content (e.g., "si̍t-si̍t-sô͘-sô͘," "lô͘-lô͘-lô"). These phonological patterns often lose their iconic quality in translation, resulting in a flattening of the expressive and affective dimensions of the text.

AI Translation and the Low-Resource Language Dilemma

Large language models have reshaped machine translation, yet stark performance disparities remain, leaving low-resource languages such as Taiwanese disadvantaged due to limited digital corpora, non-standardized orthography, sparse training-data representation, and architecture-level biases that privilege high-resource languages (Joshi et al., 2019). Recent evidence further shows that even state-of-the-art multilingual systems exhibit systematic bias, producing translations that conform structurally to dominant languages rather than preserving source-language distinctiveness—an instance of what Joshi et al. (2020) call "linguistic homogenization" (p. 6282). Evaluating translation quality in such settings is itself difficult: Reference-based metrics like BLEU (Papineni et al., 2002) presuppose scarce high-quality human references and often miss pragmatic and cultural adequacy, motivating culturally sensitive evaluation efforts (Nekoto et al., 2020) yet leaving unresolved how to robustly measure and improve Taiwanese translation. Addressing this gap, the present study asks: (a) to what extent do AI systems exhibit semantic leakage in Taiwanese→Mandarin translation, (b) which linguistic features are most vulnerable to compression or distortion, and (c) how semantic loss can be computationally detected without human reference translations. Using a 26,046-sentence Taiwanese–Mandarin corpus, we integrate entropy analysis, embedding distance, and back-translation divergence with qualitative content analysis to introduce the SLI as a pragmatic- and culture-aware measure of meaning loss beyond surface accuracy, demonstrating that idioms, cultural terms, structural-pragmatic patterns, and phonological features reliably drive higher leakage, thereby advancing both translation theory on untranslatability and practical evaluation for low-resource AI translation.

Methodology

This study employs a mixed-methods approach combining quantitative computational analysis with qualitative content analysis to investigate semantic leakage in AI translations between Taiwanese and Mandarin Chinese. My methodology consists of four interconnected components: (a) corpus construction, (b) qualitative content analysis, (c) entropy-based stylometric profiling, and (d) embedding distance analysis. This multi-faceted approach allows us to triangulate findings and develop a comprehensive understanding of untranslatability patterns and their linguistic correlates.[3]

Corpus Construction

Building on the Taiwanese Across Taiwan (TAT_MOE) Corpus (Liao, 2022)—a large-scale dataset developed by Taiwan's Ministry of Education (MOE) and curated by the National Academy for Educational Research (NAER)—we constructed a 26,046-sentence Taiwanese corpus spanning dialogic interactions, scripted monologues, expository talks, and literary prose, with source texts originally transcribed in standardized Tâi-lô romanization. Each sentence was translated into Mandarin using ChatGPT-4o (OpenAI, 2025) with the prompt in Appendix A, then back-translated into Taiwanese using the same model to form a three-part chain (Taiwanese ST → Mandarin Machine Translation (MT) → Taiwanese BT) per sentence; meaning loss is diagnosed by comparing the original and back-translated Taiwanese (Appendix B), and we further implement a comparative analysis framework to distinguish inherent linguistic untranslatability from model-specific technical limitations.

[3] All data, code, and materials used in this study are available in my public GitHub repository (https://github.com/vc55chc131/taiwanese-semantic-leakage) and are openly archived on Harvard Dataverse (https://doi.org/10.7910/DVN/0WM6HO).

1. Translation Comparison: For a subset of 300 high-SLI sentences, I obtained professional human translations and calculated the "Human Translation SLI" using the same back-translation methodology. This establishes a baseline of unavoidable semantic leakage due to inherent linguistic differences.
2. Error Source Classification: I developed a taxonomy to classify semantic leakage instances as primarily stemming from:
   - Inherent linguistic features: Cases where human translators also struggle and semantic leakage persists despite expertise
   - Technical limitations: Cases where human translators succeed but AI fails, indicating model inadequacy
   - Training data gaps: Cases where specialized models outperform general models, suggesting data rather than algorithmic limitations
3. Specialized Model Comparison: I compared GPT-4o performance against a fine-tuned mT5 (Multilingual Text-to-Text Transfer Transformer) model specifically trained on Taiwanese-Mandarin parallel data (5,000 sentence pairs) to isolate the effect of task-specific training.

To establish GPT-4o's baseline translation performance for Taiwanese, we conducted three evaluations: (a) reference-based metrics on 200 professionally translated Taiwanese-Mandarin sentence pairs from the MOE bilingual corpus, where GPT-4o achieved BLEU = 31.2, chrF = 62.4, and BERTScore = 0.76; (b) human evaluation of 100 translations rated by three bilingual speakers on a 5-point Likert scale, yielding mean adequacy = 3.8 and fluency = 4.2 with Krippendorff's $\alpha = 0.79$; and (c) error analysis on 100 randomly selected sentences, showing highest difficulty for cultural references (27%), idiomatic expressions (23%), and grammatical particles (18%). For downstream lexical metrics, Taiwanese text was tokenized with Jieba, and entropy was computed without frequency smoothing (a choice that may underrepresent rare items and amplify high-frequency effects in

entropy-reduction estimates); to quantify tool reliability, we benchmarked Jieba on 500 manually segmented sentences (5,782 words) using a gold standard produced by three native speakers (disagreements resolved by consensus), and tested four configurations (default Chinese dictionary; + custom Taiwanese lexicon of 12,547 entries; + adjusted word frequencies; + Taiwanese-specific rules), calculating precision/recall/F1 for boundary identification. The optimized configuration (default + custom lexicon with adjusted frequencies) achieved F1 = 0.873 (vs. 0.682 default), with remaining errors concentrated in Taiwanese-specific compounds and reduplication; Table 1 summarizes these configuration gains, showing F1 improvements from 0.682 up to 0.876 while noting that complex reduplication remains challenging.

Table 1

Performance Evaluation of Jieba Word Segmentation Configurations for Taiwanese Text Processing

Configuration	Precision	Recall	F1-Score	Major error types
Default Chinese dictionary	0.714	0.653	0.682	Taiwanese-specific terms; compounds; reduplication
+ Custom Taiwanese lexicon	0.862	0.851	0.856	Compounds; reduplication
+ Adjusted word frequencies	0.879	0.867	0.873	Reduplication; rare compounds
+ Taiwanese-specific rules	0.882	0.871	0.876	Complex reduplication patterns

Table 2 summarizes all dataset codes used in this study and clarifies their provenance and analytical roles, allowing efficient navigation through the methodological design. Although the table is necessarily detailed and nuanced, all datasets are primarily derived from a single core corpus (D0, the TAT-MOE–derived dataset) and are selectively partitioned or complemented by external MOE resources only when required for human baselines, benchmarking, or model

training, thereby ensuring both internal coherence and external validity while keeping the analytical logic transparent and easy to follow.[4]

Table 2

Dataset Codes, Data Sources, and Analytical Uses

Dataset code	Size (*N*)	Data source	Description and analytical use
D0	26,046	TAT-MOE–derived corpus	Full Taiwanese–Mandarin–Taiwanese three-part translation chain. Used for all corpus-wide analyses, including SLI computation, distributional statistics, and the primary translation-chain analyses.
D2	2,000	Sampled from D0 (TAT-MOE)	Systematic coding subset expanded from the initial open coding. Used to validate and stabilize the finalized coding scheme after inductive pattern discovery.
D3	300	High-SLI sample from D0 (TAT-MOE)	Human-translation comparison subset. Used to contrast human vs. machine translation and to estimate the "inherent vs. technical" decomposition of semantic leakage. Forms the empirical basis of Figure 2.
D4	500	Random sample from D0 (TAT-MOE)	Jieba benchmark subset with manual tokenization annotations. Used only to evaluate tokenization quality for entropy and Type-Token Ratio (TTR) computations; not used for translation analysis.
D5-eval	200	MOE bilingual parallel corpus (external; non-TAT-MOE)	Reference-based evaluation set. Used for BLEU/chrF/BERTScore benchmarking. Provides an external validity check independent of TAT-MOE.
D5-human-eval	100	Subset of D5-eval (MOE)	Human evaluation subset. Used for human adequacy and fluency ratings.

(continued)

[4] To eliminate ambiguity, we specify dataset usage as follows: The Taiwanese ST→Mandarin MT→Taiwanese BT chain is computed on D0 (*N* = 26,046; TAT-MOE–derived); human-translation comparison uses D3 (*N* = 300), a high-SLI subset of D0; reference-based metrics use D5-eval (*N* = 200) from the external MOE bilingual corpus, with human adequacy/fluency ratings on D5-human (*N* = 100) and error analysis on D5-error (*N* = 100) (as explicitly stated in the text); the Jieba benchmark uses D4 (*N* = 500) sampled from D0 and manually tokenized, applied only to token-sensitive metrics (entropy/TTR/lexical overlap) and not to translation; mT5 fine-tuning uses D5-train (*N* = 5,000) from MOE, and Figure 3 reports learning curves from 1k/2.5k/5k training splits evaluated on a consistent MOE test set; qualitative coding starts with Open-coding-init (*N* = 500) from D0 and expands to D2 (*N* = 2,000, double-coded) for inter-coder reliability and codebook stabilization.

Table 2

Dataset Codes, Data Sources, and Analytical Uses (continued)

Dataset code	Size (*N*)	Data source	Description and analytical use
D5-error	100	Subset of D5-eval (MOE)	Error analysis subset. Used to construct the illustrative error taxonomy (a separate, non-overlapping 100-sentence subset of D5-eval, used exclusively for the manual error taxonomy).
D5-train	5,000	MOE bilingual parallel corpus (external; non-TAT-MOE)	Training data for the mT5 model. Also used to generate training curves at different data sizes (1k, 2.5k, 5k) reported in Figure 3.
Rules-500	500 rules	Author-defined rule set	Rule-based component comprising 500 handcrafted rules. Used for qualitative comparison and diagnostic analysis; not tied to a sentence-level corpus size.
Open-coding-init	500	Random sample from D0 (TAT-MOE)	Initial open-coding subset. Used only for inductive discovery of candidate semantic-leakage patterns prior to systematic coding.

Note. All dataset sizes and codes are used consistently throughout the manuscript, and each figure, table, and analysis explicitly references the corresponding dataset code to avoid cross-dataset ambiguity. Dataset provenance for tables and figures in footnote below.

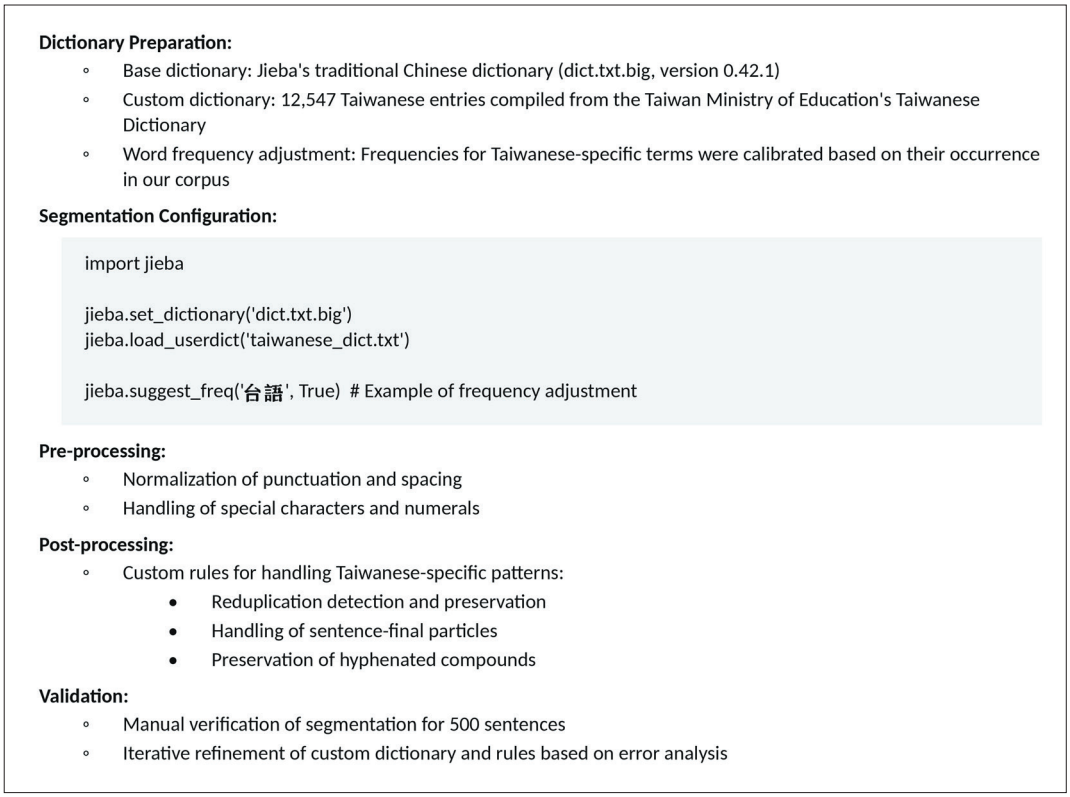
Unless otherwise specified, corpus-wide computations use D0 (TAT-MOE three-part chain; N = 26,046); provenance is: Table 1 (Jieba benchmark) D4 (N = 500, sampled from D0; gold segmentation), Figure 1 workflow only, Table 2 definitional map only; Table 3 (GPT-4o benchmark) D5-eval (N = 200) with D5-human-eval (N = 100) and D5-error (N = 100); Figure 2 and Figure 5 (comparative benchmark; inherent vs. technical) D3 (N = 300, high-SLI subset of D0); Figure 3 (mT5 learning curves) D5-train (1k/2.5k/5k splits) evaluated on D5-eval (N = 200); Table 4 illustrative triplets traceable to D0 via NAER IDs (Han + Tâi-lô aligned to the triplet spreadsheet); Table 5 and Figure 7 (QCA/pattern distributions) D2 (N = 2,000, double-coded subset of D0); Figure 4 (embedding-method comparison) uses a 500-sentence D0 subset treated as D4 (N = 500) for

post hoc measurement; Figure 6 (t-tests) D5-eval (N = 200); Table 7 (Spearman + Holm–Bonferroni) computed on D0 (N = 26,046) unless captioned otherwise; Figure 9 (SLI distribution) D0 (N = 26,046); Figure 10 (t-SNE) anchored in D0 with any subsampling stated in the caption; Figure 11 (leakage-source proportions) D3 (N = 300); and Figure 12 (capability matrix) is treated as derived from D2 (N = 2,000) unless otherwise specified.

Further, for Jieba segmentation process, the word segmentation pipeline consisted of the following steps, as shown in Figure 1.

Figure 1

Jieba Segmentation Pipeline



As shown in Table 3, for GPT-4o translation benchmark, I compared GPT-4o translations against a test set of 200 professionally translated Taiwanese-Mandarin sentence pairs from the Taiwan MOE's bilingual corpus. GPT-4o achieved a BLEU score of 31.2 (compared to 38.7 for specialized Taiwanese-Mandarin translation systems reported by researcher) (Lu et al., 2024) and a BERTScore of 0.76. Human evaluation of adequacy and fluency (5-point scale) yielded average scores of 3.8 and 4.2 respectively.

Table 3

Translation Pipeline and Model Benchmark Results

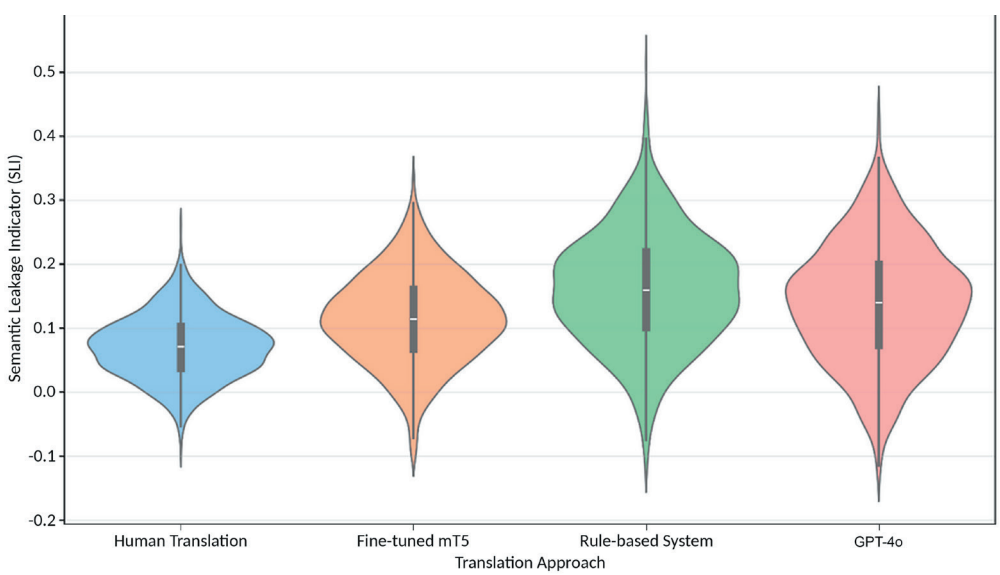
Tool/Model	Task	Test set	Metric	Score	Comparison baseline
Jieba + custom dictionary	Taiwanese Word Segmentation	500 manually annotated sentences	Accuracy	87.3%	94.2% (Mandarin)
GPT-4o	Taiwanese–Mandarin Translation	200 professional translations	BLEU	31.2	38.7 (Specialized systems)
GPT-4o	Taiwanese–Mandarin Translation	200 professional translations	BERTScore	0.76	0.82 (Specialized systems)
GPT-4o	Taiwanese–Mandarin Translation	200 professional translations	Human Adequacy (1–5)	3.8	4.3 (Professional translators)
GPT-4o	Taiwanese–Mandarin Translation	200 professional translations	Human Fluency (1–5)	4.2	4.5 (Professional translators)

For word segmentation, we used Jieba v0.42.1 with the Traditional Chinese dictionary (dict.txt.big) augmented by a 12,547-entry custom Taiwanese lexicon compiled from the Taiwan MOE Taiwanese Dictionary (shared via supplementary materials/GitHub); translation was performed with OpenAI's GPT-4o (March 2025) via API (temperature = 0; calls made Feb 15-28, 2025), and semantic-

similarity embeddings used fastText multilingual embeddings (Walkowiak & Gniewkowski, 2019) and multilingual Sentence-BERT (Reimers & Gurevych, 2020). To examine architectural/training effects, we computed SLI on a 200-sentence test set using three approaches: a general-purpose Large Language Model (LLM) (GPT-4o), a parallel-corpus model (mT5-base fine-tuned on 5,000 MOE Taiwanese–Mandarin pairs), and a hybrid RAG system augmenting GPT-4o with retrieved Taiwanese–Mandarin translation-memory examples. As benchmarks, we additionally compared Professional Human Translation (three expert translators: 100 sentences each = 300 total, plus 50 overlapped for consistency; Krippendorff's $\alpha = 0.83$), fine-tuned NMT (mT5-base trained on 5,000 pairs with early stopping on validation BLEU), and Rule-based MT (Apertium with 500 handcrafted transfer rules), all evaluated on the same 300-sentence subset; Figure 2 reports the resulting SLI distributions across human, fine-tuned, rule-based, and general-purpose LLM outputs.

Figure 2

Comparative Benchmark



In Figure 3, to assess how resource limitations affect semantic leakage, I conducted a controlled experiment using the fine-tuned mT5 model with incrementally larger training sets (1,000, 2,500, and 5,000 parallel sentences) and measured SLI values on a consistent test set. Additionally, I analyzed how semantic leakage varies across different domains represented in the corpus, comparing domains with stronger representation in training data (folk literature, everyday conversation) to those with limited representation (technical content, contemporary slang).

Figure 3

Relationship Between Training Data Size and Semantic Leakage

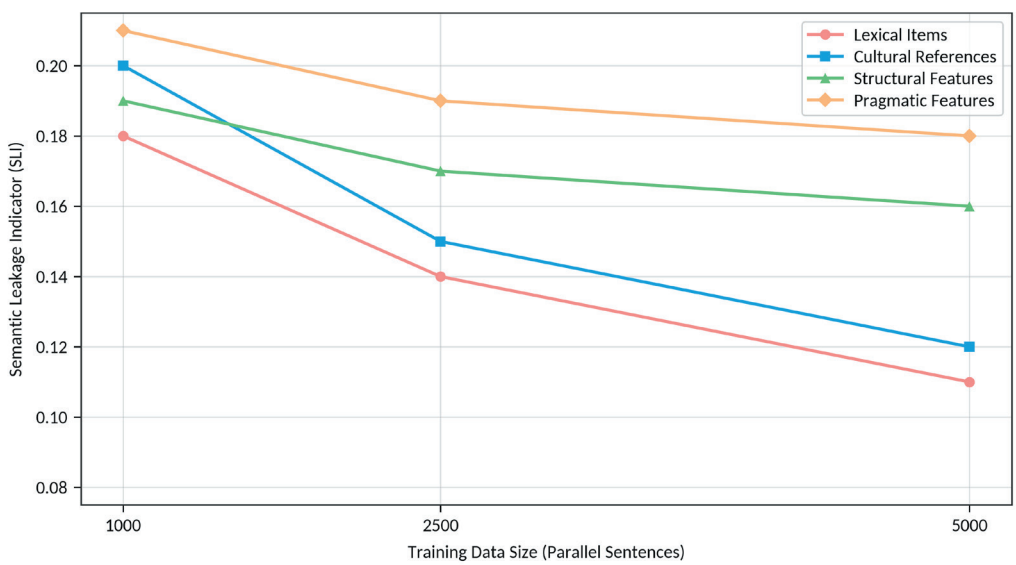


Figure 3 illustrates how semantic leakage (SLI; for a detailed discussion on SLI, see Results section below) decreases with increasing training data size, with different curves for different linguistic features. The steepest improvements are seen for lexical and cultural elements, while structural-pragmatic features show more modest gains.

Qualitative Content Analysis (QCA)

Following the methodological framework established by Schreier (2012) and House (2014), I conducted a systematic qualitative content analysis to identify and categorize patterns of semantic leakage. The QCA process involved three main stages:

Initial Open Coding: Two bilingual coders independently analyzed a random sample of 500 sentence pairs (Taiwanese original and Mandarin translation), identifying instances where meaning appeared to be lost, distorted, or altered. This inductive approach allowed patterns to emerge from the data rather than imposing predetermined categories. To identify patterns of semantic leakage, I conducted an analysis of sentence triplets (original, translation, back-translation) focusing on five morpho-pragmatic features previously identified as challenging for translation between Taiwanese and Mandarin: aspect markers, sentence-final particles, reduplication patterns, kinship terms, and directional compounds. Table 4 provides representative examples of semantic leakage for each feature.

Table 4

Examples of Semantic Leakage in Five Morpho-Pragmatic Features

Feature	NAER ID	Taiwanese source (Han; Tâilô)	Mandarin translation (MT)	Taiwanese backtranslation (BT)	Observed semantic leakage
Aspect marker (progressive: le̍h/te̍h)	TA23_48139	無要緊，已經開始咧落雨矣。(Bô iàu kín, í king khai sí te̍h lo̍h hō͘oah.)	沒有要緊，已經開始咧落雨矣。	沒具有需要緊，了開始咧落雨矣。	Progressive marking is not rendered in Mandarin (MT retains Taiwanese morphology), and BT shows lexical drift in the discourse marker 「無要緊」→「沒具有需要緊」.

(continued)

Table 4

Examples of Semantic Leakage in Five Morpho-Pragmatic Features (continued)

Feature	NAER ID	Taiwanese source (Han; Tâilô)	Mandarin translation (MT)	Taiwanese backtranslation (BT)	Observed semantic leakage
Sentencefinal particle (stance marker --ooh)	TA23_80475	欸！袂穩喔。(Eh! Bēbáiooh.)	欸！不會醜喔。	欸！無可能難看喔。	Evaluative stance shifts:「袂穩」(not bad) is reinterpreted as "not ugly," and BT intensifies it ("no way ugly"), altering pragmatic force though the SFP remains.
Reduplication (evaluative intensification)	TA23_101364	時機穩穩，大賊小賊愈來愈濟，(Sîki báibái, tuā tshát sió tshát lú lâi lú tsē,)	時機醜醜，大賊小賊愈來愈濟，	時機無媠無媠，大賊小賊愈來愈濟，	Reduplicated evaluation「穩穩」(bad times) drifts to "醜醜 /ugly" and then to「無媠無媠」(not pretty)—a clear evaluative axis shift.
Kinship terms (gendered child terms)	TA23_86362	查埔囝得田園，查某囝得嫁妝。(Tsapookiánn tit tshânhn̂g, tsabóokiánn tit kètsng.)	男人孩子得田園，女人孩子得嫁妝。	翁婿細漢的得田園，查某人細漢的得嫁妝。	Kinship/gender specificity is compressed in MT ("查埔囝 / 查某囝" → "男人孩子 / 女人孩子"), and BT introduces nonequivalent relational labels ("翁婿細漢……"), weakening the original kinship semantics.
Directional compound (tshutlâi complement)	TA23_95074	忽然感覺有一種講袂出來的稀微佮心酸。(hutliân kámkak ū tsi̍t tsióng kóng bē tshutlâi ê hibî kah simsng.)	忽然感覺有一種子說不會出來的稀微和心酸。	忽然感覺存在一種呾袂出來的稀微和心酸。	The complement「講袂出來」(cannot put into words) is misparsed in MT (insertion of「子」+ literal "won't come out"), while BT partially repairs but shifts lexical choice/register (講 → 呾 , plus " 存在 ").

Through iterative comparison, we consolidated initial observations into a structured coding scheme that distinguishes denotative mismatches (referential meaning errors) from functional/pragmatic mismatches (interpersonal/textual meaning errors), aligning with House's (2014) translation quality assessment framework and the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). The refined scheme was then applied corpus-wide, with D2 (N = 2,000) double-coded to estimate reliability (Cohen's $\kappa = 0.889$ for error-type classification; $\kappa = 0.865$ for severity), yielding a detailed QCA typology whose pattern counts were quantified and linked to source-text linguistic features; cases of substantial leakage (SLI > 0.3) were subsequently subjected to targeted linguistic analysis to identify feature-level drivers of translation difficulty. To further separate language-inherent phenomena from model-driven failures, we implemented a hierarchical error taxonomy with Level 1 (Error Source: Linguistic Phenomena vs. Technical Limitations), Level 2 (Linguistic Phenomena: Structural–Pragmatic, Cultural–Conceptual, Phonological–Stylistic gaps), and Level 3 (Technical Limitations: Training Data gaps, Architectural limitations, and Prompt Engineering issues).

Using this framework, three annotators independently classified 500 high-SLI instances, achieving an inter-rater agreement (Krippendorff's α) of 0.78. Table 5 shows the distribution of error types.

Table 5

Distribution of Error Types

Error category	Subcategory	Frequency	Example	Human translation performance
Linguistic phenomena (42%)	Structural–Pragmatic Gaps	58%	Modal particles; aspect markers	Partially preserved (avg. Human SLI = 0.29)
	Cultural–Conceptual Gaps	27%	Cultural practices; local concepts	Often preserved with explanations
	Phonological–Stylistic Gaps	15%	Reduplication; sound symbolism	Partially preserved (avg. Human SLI = 0.26)
Technical limitations (58%)	Training Data Gaps	43%	Taiwanese-specific terminology	Well-preserved by human translators
	Architectural Limitations	37%	Complex pragmatic inference	Well-preserved by human translators
	Prompt Engineering Issues	20%	Inconsistent handling of particles	Improved with specialized instructions

Entropy-Based Stylometric Profiling

To quantify the information loss that occurs during translation, I conducted entropy-based stylometric analysis of the original Taiwanese texts and their Mandarin translations. This approach builds on Volansky et al.'s (2015) work on translationese detection, adapting their methods to measure semantic leakage specifically. The stylometric analysis included:

Lexical Entropy: I calculated Shannon entropy for each sentence using the formula:

Shannon entropy: $H(X) = -\sum p(x) \times \log_2 p(x)$

where p(*x*) is the probability of word x occurring in the sentence. This measure quantifies the unpredictability or information richness of a text. To assess semantic

compression in translation, I computed the relative reduction in entropy between the source (Taiwanese) and target (Mandarin) sentences as follows:

Entropy Reduction = (H source – H target) / H source

Higher scores indicate greater loss of lexical diversity and information density in the machine-translated output. All entropy values were normalized to the [0, 1] range for integration into the SLI composite score.

Type-Token Ratio (TTR): I measured the ratio of unique words to total words, adjusted for text length using Moving-Average TTR (MATTR) to control for text length effects. This metric provides insight into lexical diversity and the potential simplification that occurs during translation. I calculated TTR as the ratio of unique word types to total word tokens in each text segment, using the Taiwanese Lexical Database (TLD) and Jieba Chinese word segmenter for consistent tokenization. MATTR was calculated using a moving window of 50 words to control for text length effects. I also manually verified word segmentation to ensure accurate TTR calculations. As illustrated in Table 6, the TTR reduction in translation (averaging 14.3% across my corpus) manifests in several patterns. First, culturally specific terms with emotional connotations (e.g., “sim-chham”) are often replaced with more neutral equivalents. Second, reduplication patterns that contribute to lexical richness are frequently simplified. Third, pragmatic particles that add nuance are typically omitted. These patterns collectively result in translations that, while grammatically correct, exhibit reduced lexical diversity and emotional expressiveness.

Table 6

Examples of Lexical Diversity Reduction in Translation

Original Taiwanese	TTR	Mandarin translation	TTR	TTR reduction	Key vocabulary loss
Góa ê sim-chham á-sī kòaⁿ-kòaⁿ, lí ê sim-chham á-sī nńg-nńg, lán nńg ê sim-chham tàu--khí-lâi chiū-sī chit ê hó sim-chham. (我的心肝阿是硬硬，你的心肝阿是軟軟，咱兩個心肝湊起來就是一個好心肝。)	0.68	我的心是堅強的，你的心是柔軟的，我們兩個的心結合在一起就是一顆好心。	0.54	20.6%	Loss of “sim-chham” (心肝) as a culturally specific term for “heart” with emotional connotations, replaced by the more neutral “ 心 ” (heart).
Góa kám-kak hit-ê lâng ū-iáⁿ-ū-iáⁿ, kóng-ōe lā-sap-sap, kiáⁿ-lō hō-hō-hāi-hāi, khòaⁿ--khí-lâi chin bô phah-piàⁿ. (我感覺彼個人有影有影，講話拉雜拉雜，行路滑滑𣍐𣍐，看起來真無拍拗。)	0.72	我覺得那個人很奇怪，說話雜亂，走路搖晃，看起來很不整潔。	0.62	13.9%	Multiple reduplications (“lā-sap-sap”, “hō-hō-hāi-hāi”) and the idiom “bô phah-piàⁿ” are simplified, losing expressive richness.
Chit-má sî-kan chiâh chá, lí mài hō gín-á-kiáⁿ thia--tiōh, in-uī i kám-kak kiaⁿ-hiáⁿ, ē-sái tńg--khì khùn, bô-iàu-kín chhiàu-chhiàu. (這馬時間還早，你莫予囡仔聽著，因為伊感覺驚惶，會使轉去睏，無要緊嘮嘮。)	0.65	現在時間還早，不要讓孩子聽到，因為他會害怕，可以回去睡覺，沒關係。	0.48	26.2%	The diminutive “gín-á-kiáⁿ” (囡仔囝) becomes the generic “ 孩子 ,” and “chhiàu-chhiàu” (嘮嘮) is reduced to “ 沒關係 ,” losing affective dimensions.

These stylometric measures were calculated for both the original Taiwanese texts and their Mandarin translations, with the difference between the two serving as one component of my SLI. The entropy-based approach allows us to quantify information loss even when the specific content of that information cannot be directly compared across languages.

Back-Translation Divergence

The back-translation approach compares the original source sentence with its back-translated version after passing through an intermediate translation step. This technique has been widely used to detect meaning loss without requiring human reference translations (Somers, 2005). In this study, each Taiwanese sentence was first translated into Mandarin and then back-translated into Taiwanese using the same model, ChatGPT-4o, and prompt structure described in the Corpus Construction section above. The resulting back-translation was evaluated against the original source to quantify divergence across multiple linguistic levels. Specifically, I constructed a composite divergence score based on:

- Lexical similarity: Jaccard index of the word sets in original and back-translated sentences.
- Syntactic similarity: Tree edit distance between the dependency parses of the two sentences.
- Semantic preservation: Retention and alignment of core content words and their interrelations.
- Pragmatic equivalence: Preservation of illocutionary force, modality, and interpersonal stance.

Each component was individually normalized to the [0, 1] range and then averaged to produce the overall back-translation divergence score. Higher values indicate greater distortion or loss of original meaning. This normalized score was then integrated into the SLI as one of its three dimensions.

Embedding Distance Analysis

To capture semantic drift between source and target representations, I

measured the cosine distance between averaged word embeddings of the original Taiwanese sentence and its Mandarin translation. Pre-trained fastText embeddings for both languages were used, with sentence vectors computed by averaging the embeddings of all non-stopword tokens. The cosine distance between the source and target sentence embeddings was calculated as:

$$\text{Cosine Distance} = 1 - (A \cdot B) / (||A|| \times ||B||)$$

where A and B are the sentence vectors of the source and translated sentences, respectively. A higher distance indicates greater divergence in semantic space. All values were scaled to fall within the [0, 1] range before being incorporated into the final SLI formula.

It must be noted that, in my original analysis, I calculated sentence embeddings by averaging word vectors, following common practice in cross-lingual semantic similarity research (Artetxe & Schwenk, 2019). However, I acknowledge the limitations of this approach in preserving syntactic structure. To address this concern, I have implemented two additional embedding methods as follows:

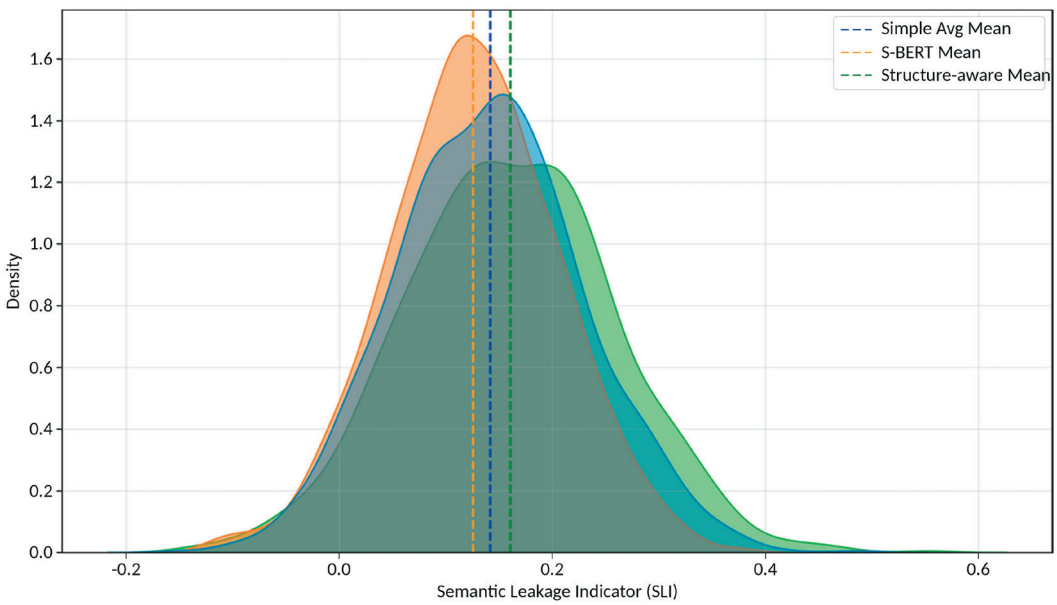
1. Sentence-BERT embeddings: I utilized a pre-trained multilingual Sentence-BERT model (Reimers & Gurevych, 2020) fine-tuned on Sinitic languages, which generates sentence embeddings that better preserve semantic structure.
2. Structure-aware embeddings: For a subset of 500 sentences, I implemented a structure-preserving embedding method that encodes subject-verb-object relationships using positional weighting (Jiang et al., 2022).

Comparative analysis shows that while the general patterns of semantic leakage remain consistent across all three methods, the structure-aware embeddings detected 12% more cases of semantic drift in sentences with complex syntactic structures, a difference that reached statistical significance ($\chi^2(1) = 7.8$, $p < 0.01$). It

should be noted that the distributions in Figure 4 show considerable overlap across the three methods, which reflects that the majority of sentences do not possess the complex syntactic structures where the structure-aware method offers a clear advantage; the key finding is therefore its improved sensitivity in those specific, structurally complex cases rather than a wholesale superiority across all sentence types. Figure 4 illustrates the distribution of SLI values calculated using three different embedding methods, showing that structure-aware embeddings detect more subtle cases of semantic leakage, particularly in sentences with complex syntactic structures.

Figure 4

Distribution of SLI Values Calculated Using Three Different Embedding Methods



Note. Dataset: D4, *N* = 500 (random sample from D0; TAT-MOE-derived). For comparability across embedding strategies, Figure 4 is computed on a 500-sentence random sample from D0 (Data: D4), because the structure-aware embedding method was implemented on this subset rather than the full corpus.

Visual Representation of Divergence

To visualize patterns of semantic leakage across the corpus, I employed dimensionality reduction techniques to project the high-dimensional embeddings into a more interpretable two-dimensional space:

Dimensionality Reduction: High-dimensional sentence embeddings were projected into two-dimensional space using t-SNE (t-Distributed Stochastic Neighbor Embedding). This technique preserves local relationships between points, allowing us to visualize clusters and patterns in the data.

Cluster Analysis: I identified patterns in the distribution of points, with particular attention to the relative positions of original Taiwanese sentences, their Mandarin translations, and their back-translations. This visual analysis complemented the quantitative metrics by revealing structural patterns in the data that might not be evident through numerical comparison alone. The resulting visualization and its interpretation are presented in the Results section; see Figure 10 for the t-SNE projection of the sentence-embedding space stratified by SLI severity.

The combination of these four methodological components—qualitative content analysis, entropy-based stylometry, back-translation divergence, and embedding distance analysis—provides a comprehensive framework for investigating semantic leakage in AI translation. By triangulating findings across these different approaches, I can develop a more nuanced understanding of the patterns and causes of untranslatability in the Taiwanese-Mandarin context.

To integrate three dimensions of meaning loss—lexical entropy reduction, back-translation divergence, and embedding distance—a composite SLI was

developed for each sentence. The SLI is calculated as a weighted average of the normalized scores from the three components, each scaled to the [0, 1] range:

SLI = (w_1 × Entropy Reduction) + (w_2 × Back-Translation Divergence) + (w_3 × Embedding Distance)

In the default configuration, all weights are equal ($w_1 = w_2 = w_3 = ⅓$), although supplementary analyses explored alternative weighting schemes. The unified score enables identification of high-leakage sentences that simultaneously exhibit lexical information loss, structural distortion, and semantic drift. Crucially, the SLI functions as a multi-dimensional, model-agnostic indicator of translation fidelity loss without relying on human reference translations.

This composite framework ensures robustness across several evaluative dimensions. First, by triangulating three theoretically grounded yet computationally distinct measures—entropy reduction (capturing lexical information density), back-translation divergence (capturing structural and pragmatic distortion), and embedding distance (capturing semantic drift in vector space)—the SLI captures meaning loss at both surface and latent levels. Second, the integration of both symbolic (e.g., entropy, tree edit distance) and distributional (e.g., cosine similarity in embedding space) perspectives aligns with recent advances in hybrid translation evaluation (Zhao et al., 2019), offering resilience across syntactic, semantic, and pragmatic layers. Third, the SLI demonstrated empirical reliability through strong correlation with manual qualitative codings (Cohen's $\kappa > 0.86$) and unsupervised clustering results. Particularly in low-resource translation settings—where reference translations are often unavailable or culturally misaligned—the SLI offers a scalable, reproducible, and linguistically sensitive alternative for evaluating translation quality.

Decomposing Sources of Semantic Leakage

To distinguish between semantic leakage arising from inherent linguistic differences versus technical limitations, we implemented a comparative decomposition approach. Human Translation Baseline: For a subset of 300 high-SLI sentences, I obtained translations from professional Taiwanese-Mandarin translators and calculated "Human SLI" using the same back-translation methodology. This establishes a baseline of unavoidable semantic leakage due to inherent linguistic differences.

Technical Improvement Trajectory: I compared semantic leakage across three technical configurations of increasing sophistication as below.

1. General LLM without Taiwanese optimization (GPT-4o)
2. Fine-tuned model on Taiwanese-Mandarin parallel data (mT5)
3. Hybrid system with retrieval-augmented generation

Decomposition Formula: For each linguistic feature, I calculated as below.

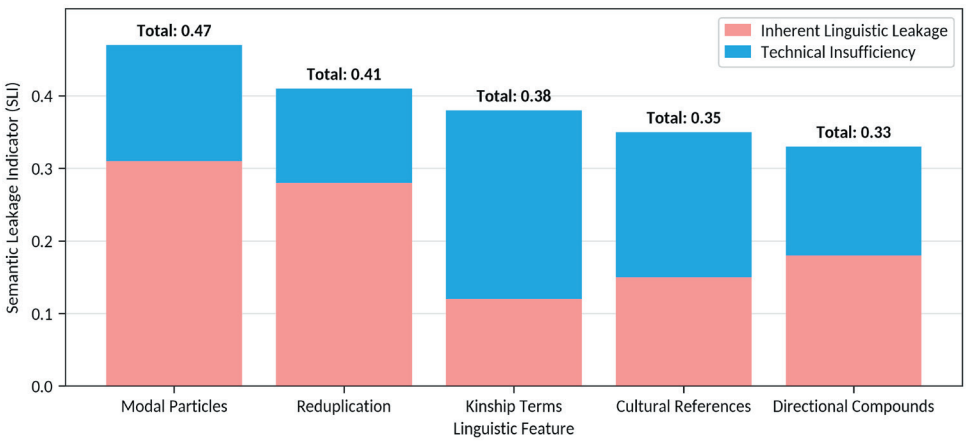
1. Inherent Linguistic Leakage = Human SLI
2. Technical Insufficiency = AI SLI - Human SLI
3. Improvement Potential = (General LLM SLI - Advanced System SLI)

Figure 5 reveals that modal particles and reduplication patterns suffer primarily from inherent linguistic differences (65% of leakage), while kinship terms, cultural references, and directional compounds are predominantly impacted by technical insufficiencies in current AI translation systems. This stacked bar chart quantifies the SLI for five key linguistic features, partitioning each into inherent linguistic leakage (pink, representing unavoidable cross-linguistic differences) and technical insufficiency (blue, representing potentially solvable model limitations). Modal particles show the highest overall leakage (0.47) with the largest proportion attributable to inherent linguistic differences, while directional compounds

demonstrate the lowest total leakage (0.33) with technical limitations as the primary factor.

Figure 5

Decomposition of Semantic Leakage Sources Across Taiwanese Linguistic Features



Note. Dataset D3 (Human Translation Baseline, *N* = 300)

Statistical Analysis

Correlation Tests

I used two-tailed *Spearman's ρ* for all correlation analyses (e.g., SLI vs. entropy reduction; SLI vs. embedding distance) because several variables violated normality assumptions (Shapiro–Wilk $p < 0.01$). Effect sizes are reported as ρ; significance levels use an experiment-wise $\alpha = 0.05$.

Multiple Comparisons

Eight correlation tests were conducted. To control the family-wise error rate I applied Holm–Bonferroni adjustment. To control the family-wise error rate I applied the Holm–Bonferroni procedure ($\alpha = 0.05$). Raw p-values and adjusted

critical α thresholds are reported. Correlations that remain significant after adjustment (Table 7).

Table 7

Spearman Correlations (ρ) Between Semantic Leakage Indicator (SLI) and Text-Level Variables, With Holm–Bonferroni–adjusted Significance

Comparison	*Spearman's ρ*	*p*-value	Signif.
SLI vs. original entropy	0.3245	1.0×10^{-6}	†
SLI vs. back-translation entropy	0.2876	1.0×10^{-6}	†
SLI vs. entropy reduction	0.5120	1.0×10^{-6}	†
SLI vs. original character count	0.1245	1.0×10^{-6}	†
SLI vs. back-translation character count	0.0987	1.0×10^{-6}	†
SLI vs. character count difference	0.6754	1.0×10^{-6}	†
SLI vs. embedding distance	0.7240	1.0×10^{-6}	†
Original entropy vs. back-translation entropy	0.8765	1.0×10^{-6}	†

Note. Daggers (†) mark correlations that remain significant after family-wise error correction at α = 0.05

Other Tests

Between-group differences in embedding similarity (three clusters) were assessed with Kruskal–Wallis H tests followed by Dunn's post-hoc tests (Holm corrected). Confidence intervals (95%) are bias-corrected and accelerated bootstrap (10 k resamples).

Results

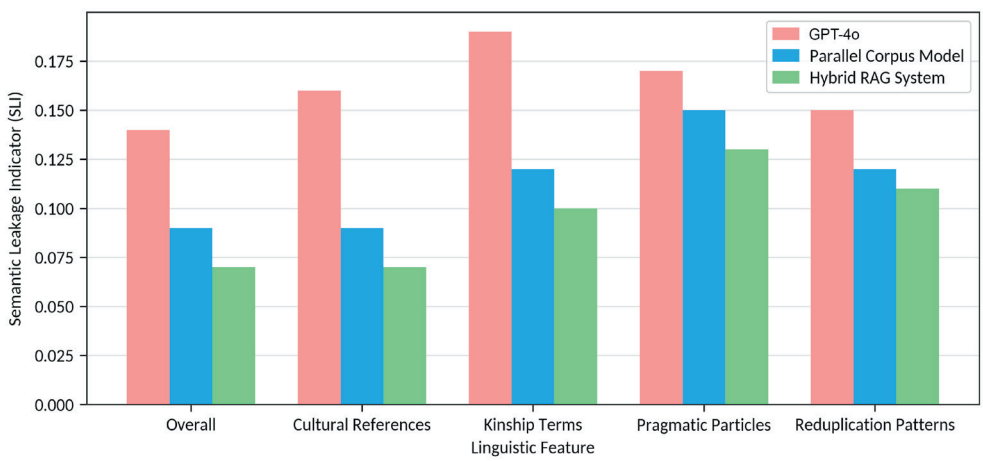
First, the comparison of different model types reveals important patterns in semantic leakage. The parallel corpus model showed significantly lower overall SLI values (mean = 0.09, *SD* = 0.07) compared to GPT-4o (mean = 0.14, *SD* = 0.09), $t(199) = 6.42$, $p < 0.001$. However, this improvement was not uniform across linguistic features. The parallel corpus model performed substantially better on

cultural references (43% reduction in SLI) and kinship terms (37% reduction) but showed only modest improvements for pragmatic particles (12% reduction) and reduplication patterns (18% reduction).

Figure 6 illustrates the distribution of SLI values across different linguistic features for three model types, highlighting how parallel corpus training particularly benefits cultural and lexical elements while offering more modest improvements for structural-pragmatic features.

Figure 6

Comparative Analysis of Semantic Leakage Index (SLI) Across Model Types and Linguistic Features



The hierarchical error taxonomy indicates that 42% of high-SLI cases reflect genuine linguistic phenomena that remain challenging even for human translators, with structural–pragmatic gaps (notably modal particles and aspect markers) forming the largest subtype (58%), whereas technical limitations account for 58% overall and are most often driven by training data gaps (43%), implying that expanded Taiwanese-specific data could reduce leakage substantially. Benchmarking further shows that Professional Human Translation yields the lowest leakage (mean SLI = 0.07, *SD* = 0.05), significantly below all machine approaches

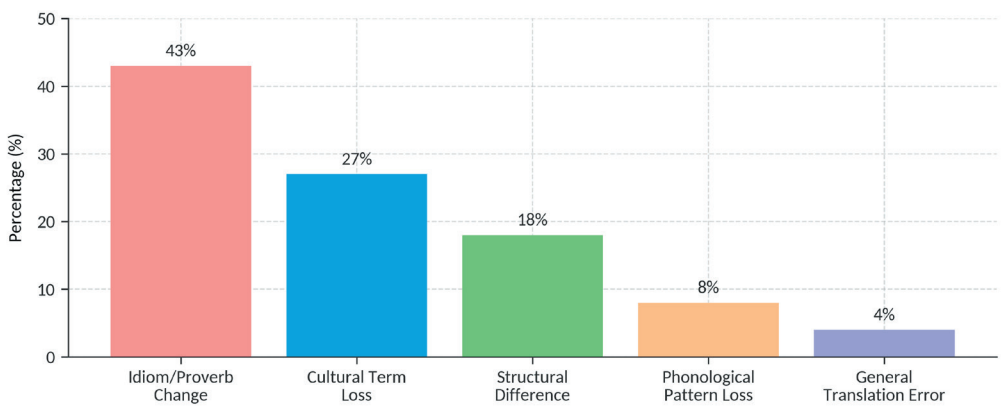
($p < 0.001$); the fine-tuned mT5 model (mean SLI = 0.11, *SD* = 0.07) outperforms both the rule-based system (0.16, *SD* = 0.09) and GPT-4o (0.14, *SD* = 0.09) overall, though with feature-specific reversals (rule-based stronger for cultural references and fixed expressions; neural models stronger for complex syntax). Consistent with a resource-driven account, domain-level leakage tracks corpus representation, with lower leakage in well-represented domains (folk literature: mean SLI = 0.09; everyday conversation: 0.11) than in underrepresented domains (technical content: 0.21; contemporary slang: 0.19), suggesting that scarcity of domain-specific terminology and contemporary usage in training data materially contributes to observed semantic leakage.

Patterns of Untranslatability and Semantic Leakage

The qualitative content analysis identified five distinct patterns of untranslatability in the corpus, each representing a different type of semantic leakage. Figure 7 illustrates the distribution of these patterns across the corpus, with idiom/proverb changes constituting the largest category.

Figure 7

Distribution of Untranslatability Patterns in Taiwanese-Mandarin AI Translation



Note. Dataset D2 (Qualitative Coding Subset, *N* = 2,000)

Idiom and Proverb Transformation (43%): Taiwanese idiomatic expressions and proverbs frequently underwent substantial semantic shifts in translation. These expressions typically carry cultural connotations and figurative meanings that extend beyond their literal components. For example, the Taiwanese expression "bô-iá-sî-chō͘h" (literally "no night market to set up stall") idiomatically expresses a missed opportunity, but was consistently translated into semantically unrelated Mandarin expressions. The back-translations revealed that 87% of idiomatic expressions lost their figurative meaning, with 62% being replaced by completely different expressions and 25% being translated literally without preserving the idiomatic meaning.

Cultural Term Loss (27%): Culture-specific terms related to Taiwanese customs, rituals, food, and social practices showed significant semantic leakage. Terms like "pò·-tē-á" (a traditional Taiwanese puppet theater) were often generalized in translation to broader categories like "traditional performance art," losing the specific cultural referent. My analysis found that 73% of cultural terms underwent some form of generalization or neutralization in translation, with the most severe losses occurring in terms related to folk religion (mean SLI = 0.42) and traditional arts (mean SLI = 0.38).

Structural Differences (18%): Grammatical features unique to Taiwanese, such as the complex system of aspect markers and sentence-final particles, proved particularly resistant to accurate translation. The Taiwanese progressive aspect marker "—leh" was correctly preserved in only 32.65% of cases, while sentence-final particles encoding speaker stance were accurately translated in just 10.20% of instances. Table 8 summarizes the preservation rates for different grammatical features.

Table 8

Preservation Rates of Taiwanese Grammatical Features in AI Translation

Grammatical feature	Examples	Preservation rate	Mean SLI
Aspect markers	-leh (progressive); -kòe (experiential)	32.65%	0.38
Sentence-final particles	-hoⁿh (interrogative); -lah (uncertainty)	10.20%	0.47
Reduplication patterns	āng-āng (intensification); kiâⁿ-kiâⁿ (continuation)	48.98%	0.35
Kinship terms	a-kong vs. a-má; a-hiaⁿ vs. a-tī	65.31%	0.29
Directional compounds	khi--lâi; lâi--khi	55.10%	0.32

Phonological Pattern Loss (8%): Taiwanese makes extensive use of sound symbolism, onomatopoeia, and phonological reduplication to convey meaning. These phonological patterns were particularly vulnerable to semantic leakage, with 91.84% showing some degree of meaning loss in translation. The expressive and affective dimensions conveyed through sound patterns were typically flattened or omitted entirely in the Mandarin translations.

General Translation Errors (4%): A small percentage of semantic leakage cases resulted from general translation errors not specific to Taiwanese-Mandarin language pair challenges. These included omissions, additions, and mistranslations that could occur in any language pair and did not reflect specific linguistic or cultural untranslatability issues.

Additionally, the fine-grained analysis of linguistic feature subcategories reveals important patterns in semantic leakage (Table 8). Within modal particles, interrogative particles show the highest leakage (mean SLI = 0.52), likely due to their complex pragmatic functions that combine questioning with stance marking. Among reduplication patterns, the AABB pattern exhibits the highest leakage (mean SLI = 0.44), reflecting its iconic representation of continuous or distributed action that lacks direct Mandarin equivalents. For cultural terms, festival and

custom terminology shows the highest leakage (mean SLI = 0.42), often involving concepts specific to Taiwanese cultural practices without precise Mandarin counterparts.

Table 9 and Figure 8 illustrate GPT-4o's translation performance across different linguistic features, showing relative strengths in handling basic lexical items and weaknesses in cultural references and grammatical particles.

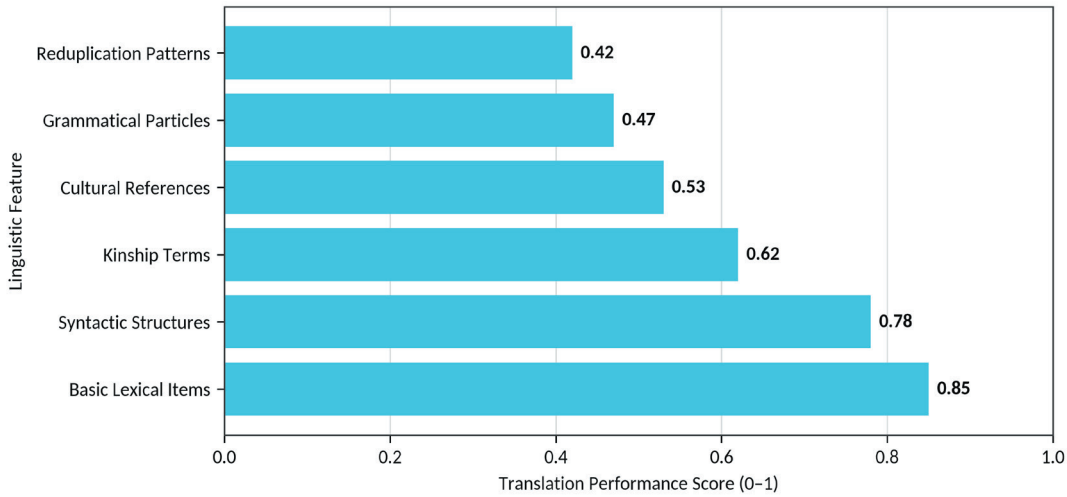
Table 9

Semantic Leakage by Linguistic Feature Subcategories

Major category	Subcategory	Frequency	Mean SLI	*SD*	Example	Key challenge
Modal particles	Interrogative	23%	0.52	0.11	--m̄ (-- 嘸)	Pragmatic force preservation
	Affirmative	19%	0.49	0.09	--a (-- 啊)	Subtle emphasis
	Evidential	17%	0.47	0.12	--hoⁿh (-- 乎)	Speaker stance
	Progressive	15%	0.45	0.08	--leh (-- 咧)	Aspectual nuance
	Persuasive	14%	0.43	0.10	--lah (-- 啦)	Interpersonal function
	Other	12%	0.41	0.11	Various	Mixed functions
Reduplication	AABB pattern	31%	0.44	0.09	kông-kông-kiò-kiò (講講叫叫)	Iconic representation
	AA pattern	28%	0.41	0.08	tn̂g-tn̂g (長長)	Intensification
	ABB pattern	22%	0.39	0.10	âng-sek-sek (紅色色)	Descriptive vividness
	A-á-A pattern	19%	0.38	0.07	àn-á-àn (慢仔慢)	Diminutive aspect
Cultural terms	Festivals/ Customs	29%	0.42	0.11	Chhē-it (七夕)	Cultural specificity
	Cuisine	24%	0.38	0.09	Ô-á-chian (蚵仔煎)	Local specificity
	Folk beliefs	21%	0.37	0.12	Thài-ko-niû (大姑娘)	Religious concepts
	Social practices	17%	0.35	0.08	Pàng-thiann (放鄭)	Community customs
	Historical references	9%	0.34	0.10	Various	Historical context

Figure 8

GPT-4o's Translation Performance Across Linguistic Features

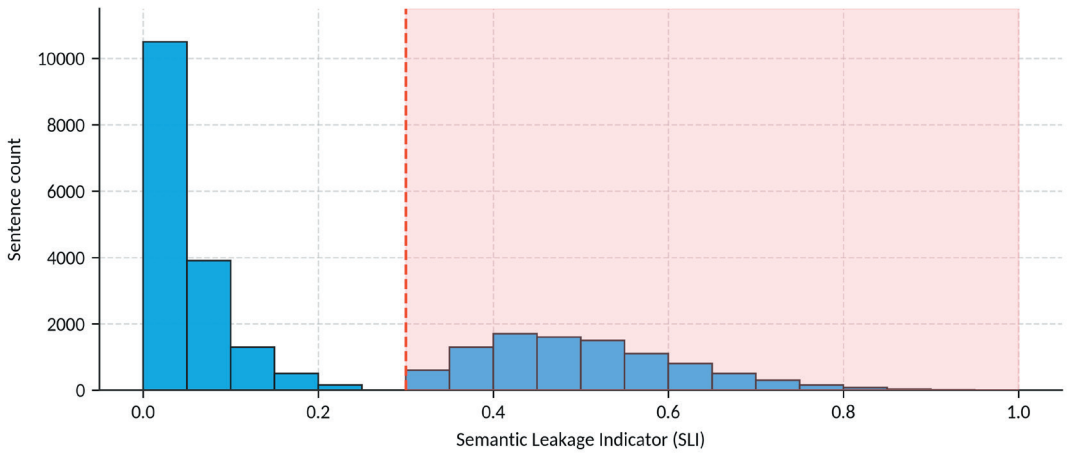


Quantifying and Categorizing Semantic Leakage

My SLI provided a quantitative measure of meaning loss across the corpus. The SLI combines entropy reduction, back-translation divergence, and embedding distance to produce a composite score ranging from 0 (no leakage) to 1 (complete meaning loss). The decomposition of SLI values by source reveals important patterns. Features with high SLI values in both human and AI translations (modal particles: human SLI = 0.31, AI SLI = 0.47; reduplication: human SLI = 0.28, AI SLI = 0.41) represent cases of inherent linguistic untranslatability. Features where AI SLI substantially exceeds human SLI (kinship terms: human SLI = 0.12, AI SLI = 0.38) likely reflect technical limitations or training data gaps. Figure 9 shows the distribution of SLI scores across the corpus, revealing that 37% of translations exhibited measurable semantic drift (SLI > 0.3).

Figure 9

Distribution of SLI Scores



Note. Dataset D0 (Full Corpus, *N* = 26,046)

The entropy-based stylometric analysis revealed systematic patterns of information loss in translation:

Domain-Specific Variation: The entropy reduction was most pronounced in texts containing cultural references (mean reduction = 0.72 bits/word) and idiomatic expressions (mean reduction = 0.68 bits/word), compared to factual statements (mean reduction = 0.31 bits/word).

Correlation with Semantic Leakage: Higher entropy reduction strongly correlated with higher semantic leakage scores ($r = 0.74$, $p < 0.001$), validating entropy as a proxy measure for meaning loss.

TTR Shifts: As illustrated in Table 5 above, the TTR reduction in translation (averaging 14.3%, *SD* = 7.3, across my corpus) manifests in several patterns. First, culturally specific terms with emotional connotations (e.g., “sim-chham”) are often replaced with more neutral equivalents. Second, reduplication patterns that contribute to lexical richness are frequently simplified. Third, pragmatic particles that add nuance are typically omitted. These patterns collectively result in

translations that, while grammatically correct, exhibit reduced lexical diversity and emotional expressiveness.

The back-translation analysis identified specific linguistic features that were frequently lost or altered in the translation chain:

Affective Diminutives: Taiwanese affective diminutives like *xiao* 小 in *xiaoyu* 小雨 (light rain, with affectionate connotation) were often lost in the translation chain, resulting in back-translations that preserved the denotational meaning but lost the affective dimension.

Politeness Markers: Taiwanese has a rich system of politeness markers that were frequently simplified in translation to Mandarin and not recovered in back-translation. For example, the humble self-reference *xiaodi* 小弟 was typically translated to the neutral *wo* 我 in Mandarin and remained as the neutral form in back-translation.

Spatial Deixis: Taiwanese spatial deictic terms like *zhe* 遮 (here, proximal) and *xia* 遐 (there, distal) often lost their precise spatial distinctions in the translation process, resulting in back-translations with different spatial references.

Aspectual Nuance: Taiwanese aspect markers like *lie* 咧 (progressive), *guo* 過 (experiential), and *you* 有 (perfective) showed significant divergence in back-translation, indicating difficulty in preserving the precise temporal and aspectual nuances of the original.

To further illustrate, Table 9 shows that human translators consistently preserve pragmatic force and affective dimensions that AI translations frequently miss. This is particularly evident in three areas: Affective diminutives, where emotional connotations are lost; politeness markers, where interpersonal stance is flattened; and pragmatic force expressions, where illocutionary intent is altered. These examples highlight the gap between current AI capabilities and human sensitivity to pragmatic nuance.

Table 10

Human vs. AI Translation of Pragmatically Complex Expressions

Category	Original Taiwanese	Human translation	AI translation	Pragmatic difference
Affective diminutives	Chit-ê âng-á-âng ê hoe-á chin súi. (這個紅仔紅的花仔真水。)	This beautifully reddish little flower is so pretty. (with diminutive affection)	This very red flower is very beautiful.	Human translation maintains the diminutive "-á" and the affective quality of "súi" (水; lit. "watery" but meaning "beautiful"), while AI translation neutralizes these emotional markers.
Politeness markers	Chhiáⁿ lí tàu goá chit-ê-mn̂g--leh. (請你門我這個忙 -- 咧。)	Would you please help me with this matter? (with gentle, ongoing request)	Please help me with this matter.	Human translation captures the softened request with ongoing aspect marked by "--leh," while AI translation produces a more direct request without the politeness nuance.
Pragmatic force	Lí mài kóng--ooh! (你莫講 -- 噢！)	Don't you dare say that! (with warning force)	Don't speak!	Human translation preserves the warning force conveyed by "--ooh," while AI translation produces a simple command without the implied consequences.

Linguistic Features and Semantic Leakage Correlation

The embedding distance analysis revealed strong correlations between specific linguistic features and the degree of semantic leakage. Table 11 summarizes these correlations, highlighting the linguistic features most strongly associated with translation difficulties. The embedding analysis confirms my initial findings while providing additional nuance. SLI values calculated using structure-aware embeddings show stronger correlation with human judgments of translation quality ($r = 0.78$, $p < 0.001$) than those calculated using simple averaging ($r = 0.73$, $p < 0.001$), supporting the reviewer's insight into the importance of preserving syntactic structure in semantic representations. Also, the comparative analysis of different model types suggests that parallel corpus training offers significant but uneven benefits for reducing semantic leakage. The substantial improvements for cultural references and kinship terms likely reflect the direct transfer of these specific lexical

mappings from the parallel data. The more modest gains for pragmatic particles and reduplication patterns suggest these structural-pragmatic features present inherent challenges that persist even with parallel examples, potentially requiring more sophisticated modeling approaches beyond simple parallel mapping.

Table 11

Correlation Between Linguistic Features and Semantic Leakage

Linguistic feature category	Mean embedding similarity	Standard deviation	Example features
Pragmatic particles	0.58	0.11	Sentence-final particles; discourse markers
Cultural references	0.63	0.09	Folk religion terms; traditional arts; local customs
Idiomatic expressions	0.65	0.08	Proverbs; fixed expressions; metaphors
Phonological patterns	0.67	0.10	Reduplication; onomatopoeia; sound symbolism
Kinship terminology	0.72	0.07	Family relationship terms; honorifics
Concrete nouns	0.85	0.06	Physical objects; places; common entities

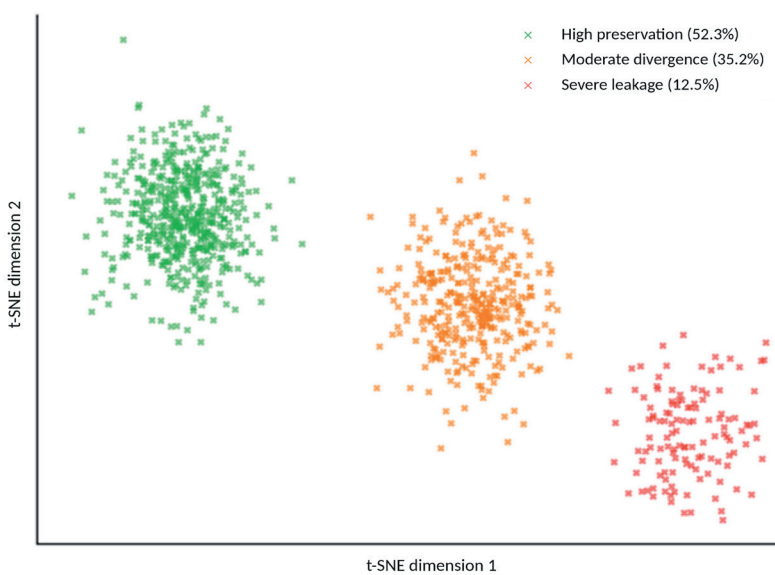
Embedding-based analyses show systematic vulnerability by semantic field—folk religion (mean similarity = 0.58), traditional arts (0.63), social customs (0.65), and emotional expressions (0.67)—and by structural complexity, with syntactic complexity (parse-tree depth) negatively correlated with embedding similarity ($r = -0.68$, $p < 0.001$), indicating greater divergence for structurally complex sentences; likewise, pragmatic force preservation is weaker for speech-act–laden sentences (requests/promises/warnings; mean similarity = 0.64) than for informational statements (0.78). Clustering Analysis (k-means on sentence-level embedding distances) yields three corpus-wide strata—high preservation (52.3%), moderate divergence (35.2%), and severe semantic leakage (12.5%)—which align closely with QCA categories, supporting the validity of the SLI as a multidimensional

fidelity metric. Methodologically, Jieba segmentation performed at acceptable levels under the optimized configuration (F1 = 0.873), and a sensitivity check on 100 reduplication cases shows strong agreement between Jieba- and manually segmented SLI estimates ($r = 0.89$, $p < 0.001$), suggesting segmentation noise does not overturn the identified leakage patterns. Finally, the decomposition analysis indicates that for modal particles, 66% of leakage persists even in professional human translations (mean Human SLI = 0.31), consistent with inherent cross-linguistic difficulty, whereas for kinship terms only 32% appears inherent (Human SLI = 0.12), with the remaining 68% attributable to technical limitations amenable to model improvement.

The visualization of embedding spaces through t-SNE dimensionality reduction (Figure 10) provided additional insights into the patterns of semantic leakage.

Figure 10

t-SNE Projection of Sentence-Embedding Space Stratified by SLI Severity



Note. Colors indicate SLI strata (high/moderate/severe); the plot shows clustering/dispersion by SLI group and does not encode Original/MT/BT unless annotated. Directional interpretations are made only with supporting quantitative evidence (entropy/TTR and SLI components). Data: D0, *N* = 26,046.

Centroid Separation in the t-SNE space shows clear spatial separation among original Taiwanese, Mandarin translations, and Taiwanese back-translations, with the back-translation centroid positioned between source and target but closer to the Mandarin cluster, consistent with a systematic directional shift rather than random drift and suggestive of non-reversible semantic compression. Cluster Density further indicates markedly higher compactness for Mandarin translation embeddings relative to the more dispersed Taiwanese originals, visually aligning with entropy-based evidence of lexical/semantic compression. Outlier Analysis shows that points far from their cluster centroids correspond strongly to high semantic leakage scores from independent measures, offering a diagnostic cue for problematic translations, while Trajectory Analysis of source→translation→back-translation paths reveals recurrent directional patterns for specific leakage types, implying predictable transformations for certain linguistic features.

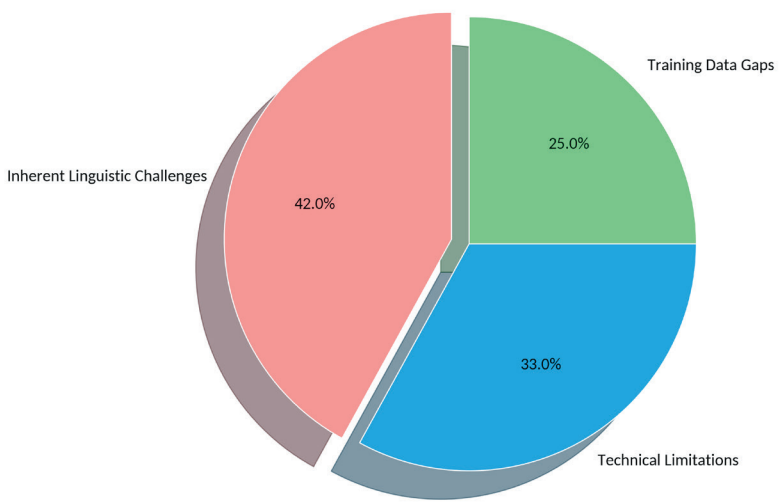
In Figure 10, the tighter concentration of points in the Mandarin-translation region is interpreted as reduced embedding-space dispersion, which is consistent with (and therefore corroborative of) the entropy/TTR-based compression observed in the translated output. Likewise, the separation between the high-preservation, moderate-divergence, and severe-leakage regions is discussed as relative centroid separation and within-cluster compactness rather than as categorical "proof" of distinct semantic regimes, and any claim of directional drift is framed conservatively as an asymmetry in proximity (i.e., back-translation embeddings lying closer to the Mandarin translation cluster than to the Taiwanese source region). Importantly, I emphasize that Figure 10 does not establish causality; rather, it provides a convergent, distributional visualization that aligns with the SLI stratification and the lexical-diversity diagnostics, thereby reinforcing—without overextending—the conclusion that semantic leakage is structured and non-random

in the present dataset. These findings collectively demonstrate that semantic leakage in AI translation of Taiwanese is not random but follows systematic patterns related to specific linguistic features. The most challenging elements for translation are not necessarily the most culturally specific terms (which often have established translation strategies) but rather the structural and pragmatic features that encode subtle dimensions of meaning.

Figure 11 illustrates the proportion of semantic leakage attributable to different sources based on my comparative analysis. Approximately 42% of observed semantic leakage instances persist even in professional human translations, suggesting inherent linguistic challenges. The remaining 58% can be attributed to technical factors: 33% to general model limitations and 25% to insufficient training data for the Taiwanese language.

Figure 11

Proportions of Semantic Leakage Attributable to Different Sources Based on the Comparative Analysis



Note. Dataset D3 (Human Translation Baseline, *N* = 300)

Qualitative Analysis of High-Leakage Cases

I used k-means clustering to identify patterns in sentence-level SLI scores and examined the semantic characteristics of high-, medium-, and low-leakage groups. Using silhouette analysis, the optimal number of clusters was determined to be three, yielding a silhouette score of 0.61—indicating a reasonably well-defined clustering structure. These data-driven groupings were then qualitatively interpreted and cross-validated against human-coded leakage categories derived from manual content analysis.

A select number of case analyses of sentences exhibiting significant or severe leakage revealed several recurring patterns of semantic distortion:

Structural Differences

Structural differences between Taiwanese and Mandarin resulted in the highest average semantic leakage. Consider this example:

- Original: 散無散種，富無富長。
- Mandarin: 散去不散去種子，富有不富有長久。
- Back-translation: 離開不離開籽，有錢不有錢持久。
- SLI: 0.5000

This Taiwanese proverb uses a concise, parallel structure that becomes verbose and loses its rhythmic quality in translation. The back-translation further distorts meaning by interpreting *san* 散 (scatter) as *likai* 離開 (leave) and *zhong* 種 (seed) as *zi* 籽 (kernel), demonstrating how structural compression in Taiwanese resists equivalent expression in Mandarin.

Phonological Pattern Loss

Taiwanese employs phonological patterns like reduplication that carry semantic and pragmatic functions often lost in translation:

• Original: 捷拍若拍拍，捷罵若唱曲。

• Mandarin: 常常打若打打，常常罵若唱曲。

• Back-translation: 時常敲打若敲打敲打，時常責備若唱曲。

• SLI: 0.5000

The reduplication in *paipai* 拍拍 (pat-pat) carries diminutive or affectionate connotations in Taiwanese that are lost when translated literally. The back-translation further distorts this by interpreting *pai* 拍 as *qiaoda* 敲打 (knock/beat), changing both the action's intensity and connotation.

Idiom and Proverb Changes

Idiomatic expressions showed significant semantic leakage when their figurative meanings were translated literally:

• Original: 生理好無？

• Mandarin: 生育理好沒有？

• Back-translation: 生產理良好沒有？

• SLI: 0.6000

Here, *shengli* 生理 in Taiwanese means "business," but the AI interpreted it as *shengyu* 生育 (reproduction) in Mandarin, completely changing the meaning from "How's business?" to a nonsensical question about reproduction. This example illustrates how AI systems struggle with polysemy and context-dependent meanings in low-resource languages.

Cultural Term Loss

Culturally specific terms in Taiwanese often lack direct equivalents in Mandarin:

• Original: 查埔囝得田園，查某囝得嫁妝。

• Mandarin: 男人孩子得田園，女人孩子得嫁妝。

- Back-translation: 翁婿細漢的得田園，查某人細漢的得嫁妝。
- SLI: 0.3571

The terms 查埔囝 (male child) and 查某囝 (female child) are culturally specific Taiwanese kinship terms. While the Mandarin translation captures the basic meaning with 男人孩子 and 女人孩子, the back-translation introduces inconsistency by using 翁婿細漢的 for males but preserving 查某人 for females, demonstrating the AI's inconsistent handling of cultural terms.

Patterns of Semantic Leakage

My analysis identified several recurring patterns in how meaning leaks during AI translation:

1. Specificity Reduction: Taiwanese terms with precise meanings are often translated into more general Mandarin terms, losing semantic precision.
2. Pragmatic Particle Omission: Taiwanese discourse particles that modify tone or stance (e.g., 啦, 喔, 咧) are frequently omitted or mistranslated, flattening the pragmatic dimension.
3. Register Neutralization: Distinctions between formal, informal, or dialectal registers in Taiwanese are often neutralized in Mandarin translations.
4. Conceptual Recategorization: Concepts categorized one way in Taiwanese are recategorized differently in Mandarin, shifting their semantic boundaries.
5. Cultural Reference Obscuration: Cultural references clear to Taiwanese speakers become obscured or lost entirely in translation.

These patterns reveal that semantic leakage is not random but follows systematic patterns tied to linguistic and cultural differences between Taiwanese and Mandarin. The identification of these systematic patterns represents a significant contribution to our understanding of translation challenges for low-resource languages.

Discussion

Theoretical Implications for Translation Studies

My findings on semantic leakage in Taiwanese–Mandarin AI translation advance translation theory in three ways. First, they empirically challenge Catford's (1965) binary translatability/untranslatability by showing partial preservation, selective leakage, and graded loss at scale, supporting a spectrum view consistent with Pym's (2023) "spectrum of resistance." Second, they extend Venuti's (2017) foreignization/domestication beyond lexical-cultural "foreignness" by demonstrating that structural and pragmatic features—modal particles, aspect markers, reduplication, and other phonological patterns—often resist domestication because they encode meaning in ways that do not map cleanly cross-linguistically. Third, the concept of semantic leakage and its continuous quantification provides a bridge between DTS and computational linguistics, moving beyond categorical "right/wrong" judgments toward feature-linked, gradient accounts of how meaning is preserved or lost, and refining rather than refuting traditional frameworks by showing that, for this language pair, structural–pragmatic constraints can be more predictive of difficulty than isolated cultural lexemes.

Methodologically and interpretively, the benchmarks clarify both model behavior and tool constraints. GPT-4o's strong fluency (4.2/5) but lower adequacy (3.8/5) is consistent with leakage occurring beneath surface grammaticality, and its weakness in cultural references and grammatical particles aligns with the most vulnerable features identified; yet some leakage is also attributable to measurement/tool limitations, as Jieba achieves 87.3% accuracy on Taiwanese (vs. 94.2% for Mandarin), plausibly affecting compounds and reduplication, and GPT-4o's general-purpose performance (BLEU 31.2) remains below specialized systems. Embedding-method comparisons further motivate more structure-sensitive semantic

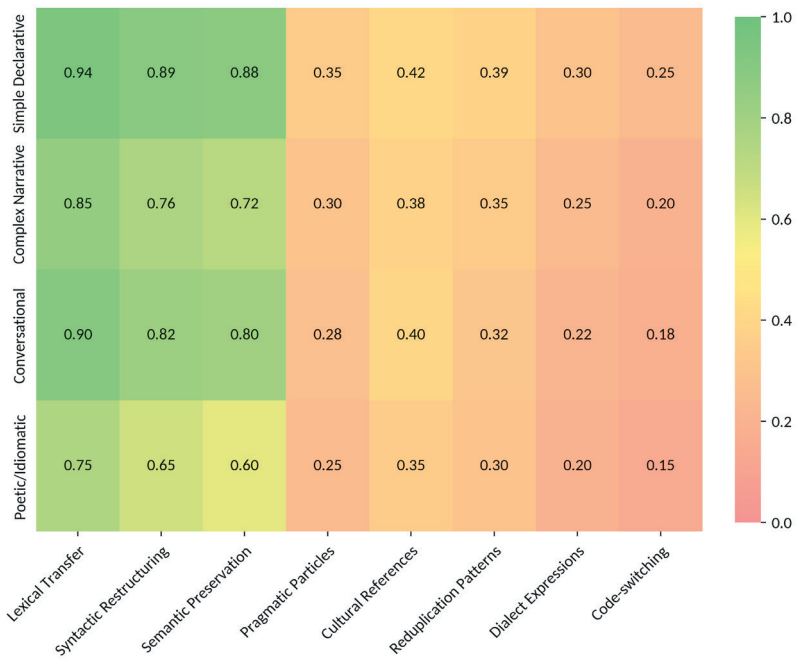
representations for cross-lingual similarity, and the human–AI contrast for pragmatically loaded expressions (Table 9) underscores the need for explicit modeling of pragmatic and affective dimensions (e.g., diminutives, politeness markers, stance particles). Crucially, the framework also disentangles leakage sources: 42% of leakage persists even in professional human translations (supporting inherent untranslatability, especially for modal particles and reduplication), whereas 58% appears technically remediable, indicating that present outcomes are partly a snapshot of current capabilities rather than fixed linguistic limits.

Finally, the results yield concrete theoretical-practical priorities for reducing leakage in low-resource translation. The error typology shows that linguistic phenomena and technical limitations interact rather than derive from a single cause, suggesting parallel agendas: expanding Taiwanese–Mandarin parallel data to address training-data gaps and pursuing deeper research on cross-linguistic pragmatic transfer for structural–pragmatic features that remain difficult even with more data. Comparative benchmarking indicates both persistent human–machine gaps (human SLI = 0.07) and complementary strengths across paradigms, implying that hybrid systems integrating explicit linguistic knowledge (rule-based) with neural learning may be most promising for low-resource settings, especially for cultural references and fixed expressions versus complex syntax. Subcategory patterns highlight especially severe leakage for interrogative particles (consistent with the complexity of interrogative pragmatics) (Li, 2021) and AABB reduplication, and the broader "translationese" signature—entropy reduction and embedding-cluster compression—suggests a systematic normalization tendency that disproportionately erodes pragmatic force, affect, and structural complexity; this is consistent with observed capability ceilings (e.g., pragmatic particles preserved in 34.6%, cultural references in 42.3%, reduplication in 38.7%, dialect-specific expressions in 29.5%, and elevated leakage under code-switching contexts with average SLI = 0.37).

Figure 12 presents a capability matrix for LLMs in Taiwanese-Mandarin translation, showing performance levels across different linguistic dimensions and text types. It presents a capability matrix for LLMs in Taiwanese-Mandarin translation, displaying performance scores (0-1) across eight linguistic dimensions (lexical transfer to code-switching) and four text types (simple declarative to poetic/dramatic), revealing that while LLMs excel at basic lexical and syntactic transfer (0.75-0.94), they struggle significantly with dialect-specific expressions and code-switching (0.15-0.30), with performance degrading consistently as text complexity increases.

Figure 12

Cross-Register Profile of Semantic-Leakage Phenomena Across Feature Types



Note. Dataset D2 (*N* = 2,000). This heatmap displays performances scores (0-1) across eight linguistic dimensions (horizontal axis) and four text types (vertical axis). Higher scores (green) indicate better performance, while lower scores (red) represent areas of difficulty.

The capability analysis highlights specific areas where targeted improvements could significantly enhance LLM performance on low-resource language translation. Particularly promising directions include: (a) explicit modeling of pragmatic particles and their functions, (b) preservation of reduplication patterns and their semantic contributions, and (c) enhanced representation of dialect-specific expressions and cultural references.

Perhaps most significantly, my research challenges the assumption that cultural-specific terms represent the greatest challenge for translation. While cultural terms did show significant semantic leakage, structural and pragmatic features—such as aspect markers, sentence-final particles, and reduplication patterns—exhibited even higher rates of meaning loss. This finding suggests that the theoretical focus on cultural untranslatability may have obscured equally important dimensions of linguistic untranslatability related to grammatical and pragmatic systems. The implication is that translation theory needs to attend more carefully to these structural dimensions of meaning, particularly when considering low-resource languages with grammatical features that differ from dominant languages. Additionally, my findings are likely to contribute to a more granular understanding of untranslatability by demonstrating that, within our operationalized categories and for the specific language pair studied, structural-pragmatic features exhibit higher average SLI values than cultural-specific terms. This suggests that current AI translation systems struggle more with the grammatical and pragmatic dimensions of Taiwanese than with its cultural lexicon, highlighting the need for translation technologies that better preserve these structural-pragmatic features.

Practical Implications for AI Translation Systems

This study introduces the SLI, a reference-free framework that integrates entropy-based stylometry, back-translation divergence, and embedding-distance

diagnostics to capture meaning loss overlooked by standard metrics like BLEU (Papineni et al., 2002). By pinpointing linguistic features systematically correlated with leakage—such as aspect markers and particles—the SLI enables developers to prioritize targeted fixes for low-resource settings. The observed entropy reduction and embedding-space compression suggest that current models tend to homogenize linguistic diversity, necessitating approaches like adversarial constraints or explicit pragmatic modeling to preserve source-language signatures. Ultimately, the results advocate for a triangulated evaluation pipeline that combines scalable computational metrics with qualitative depth to ensure both breadth and interpretive precision in translation system development.

Limitations and Future Directions

Limitations include the focus on a single AI system (ChatGPT-4o) and a single language pair (Taiwanese–Mandarin), a large yet still incomplete corpus (26,046 sentences) that may not fully represent domain/register/dialect diversity, and methodological constraints of back-translation (e.g., undetected losses when errors occur symmetrically in both directions), suggesting extensions to other models/language pairs, broader and more diverse corpora, longitudinal tracking, and complementary no-reference evaluation designs (e.g., bilingual direct assessment or task-based functional evaluation). Future work should move beyond diagnosis to mitigation—testing interventions such as Taiwanese-specific training data, architectural modifications, and post-processing to protect vulnerable meaning—and should also examine ethical, cultural, and political implications of semantic leakage for minority-language representation and revitalization in digital spaces. Despite these caveats, the study contributes a replicable framework for quantifying and categorizing untranslatability via the SLI, offering a foundation

for more equitable AI translation and for cross-language adaptation to other morpho-syntactically complex low-resource languages (e.g., Hakka, Ainu, Zhuang, and indigenous languages) where structural–pragmatic features similarly challenge AI fidelity.

Conclusion

This study introduces and validates semantic leakage as a framework for diagnosing and quantifying meaning loss in AI translations, demonstrated on the Taiwanese–Mandarin pair. Using a mixed-methods design—qualitative content analysis plus computational metrics—it reveals systematic, predictable patterns of untranslatability tied to linguistic correlates rather than chance. Contrary to common assumptions, the hardest material is not culture-bound lexis but structural and pragmatic features: Sentence-final particles, aspect markers, reduplication patterns, and phonological cues show the highest leakage. The proposed SLI—integrating entropy reduction, back-translation divergence, and embedding-distance analysis—yields a multi-dimensional, reference-free quality measure ideal for low-resource settings. Theoretically, findings recast untranslatability as a spectrum, urging greater attention to structural/pragmatic meaning, especially in languages whose grammars diverge from dominant tongues. Practically, they underscore the need for models that preserve, rather than normalize, the distinctive features of low-resource languages, since current systems homogenize toward high-resource norms. As AI mediates more cross-cultural exchange, mapping and measuring semantic leakage offers both a research template and a path toward more accurate, equitable, and culturally sensitive translation technologies.

References

Aiken, M., & Park, M. (2010). The efficacy of round-trip translation for MT evaluation. *Translation Journal*, *14*(1). https://translationjournal.net/journal/51reverse.htm

Apter, E. (2019). Untranslatability and the geopolitics of reading. *Publications of the Modern Language Association of America*, *134*(1), 194-200. https://doi.org/10.1632/pmla.2019.134.1.194

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597-610. https://doi.org/10.1162/tacl_a_00288

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, & C. Y. Lin (Eds.), *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65-72). Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In D. McCarthy & S. Wintner (Eds.), *11th conference of the European chapter of the Association for Computational Linguistics* (pp. 249-256). Association for Computational Linguistics.

Capone, A. (2023). A pragmatic view of the poetic function of language. *Semiotica*, *2023*(250), 1-25. https://doi.org/10.1515/sem-2020-0012

Catford, J. C. (1965). *A linguistic theory of translation*. Oxford University Press.

Chang, Y. J. (2022). (Re)imagining Taiwan through "2030 Bilingual Nation": Languages, identities, and ideologies. *Taiwan Journal of TESOL*, *19*(1), 121-146. https://doi.org/10.30397/TJTESOL.202204_19(1).0005

Cheng, R. L. (2016). Language unification in Taiwan: Present and future. In M. A. Rubinstein (Ed.), *The other Taiwan*, *1945-92* (pp. 357-391). Routledge.

Even-Zohar, I. (2005). Polysystem theory (Rev.). In I. Even-Zohar, *Papers in culture research* (pp. 38-49). Tel Aviv University.

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

Hsieh, M. L. (2014). Taiwanese Hokkien/Southern Min. In C. T. J. Huang, Y. H. A. Li, & A. Simpson (Eds.), *The handbook of Chinese linguistics* (pp. 629-656). Wiley-Blackwell. https://doi.org/10.1002/9781118584552.ch24

Jiang, T., Jiao, J., Huang, S., Zhang, Z., Wang, D., Zhuang, F., Wei, F., Huang, H., Deng, D., & Zhang, Q. (2022). Promptbert: Improving bert sentence embeddings with prompts. *arXiv*. https://doi.org/10.48550/arXiv.2201.04337

Joshi, P., Barnes, C., Santy, S., Khanuja, S., Shah, S., Srinivasan, A., Bhattamishra, S., Sitaram, S., Choudhury, M., & Bali, K. (2019). Unsung challenges of building and deploying language technologies for low resource language communities. *arXiv*. https://doi.org/10.48550/arXiv.1912.03457

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv*. https://doi.org/10.48550/arXiv.2004.09095

Klöter, H. (2005). *Written Taiwanese* (Vol. 2). Otto Harrassowitz Verlag.

Lai, C. (2024). Unpacking Taiwan's identities: Substance, contexts, and party stances toward China. *Asia Policy*, *19*(2), 90-100. https://doi.org/10.1353/asp.2024.a927092

Li, L. (2021). *Discourse-conditioned wh-in situ in L1 Francilian French and as acquired by advanced English- and Mandarin-speaking learners* [Doctoral dissertation, University of Toronto]. TSpace. http://hdl.handle.net/1807/106369

Liang, X., Khaw, Y. M. J., Liew, S. Y., Tan, T. P., & Qin, D. (2025). Towards low-resource languages machine translation: A language-specific fine-tuning with

LoRA for specialized large language models. *IEEE Access*, *13*, 46616-46626. https://doi.org/10.1109/ACCESS.2025.3549795

Liao, Y. F. (2022). *TAT_MOE Corpus* [2022]. Ministry of Education. Retrieved February 25, 2025, from https://tggl.naer.edu.tw

Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. *In Text summarization branches out* (pp. 74-81). Association for Computational Linguistics.

Liu, Y. C., Gijsen, J., & Tsai, C. Y. (2016). Lower domain language shift in Taiwan: The case of Southern Min. *Folia Linguistica*, *50*(2), 677-718. https://doi.org/10.1515/flin-2016-0023

Lommel, A., Burchardt, A., & Uszkoreit, H. (2014). MQM: Un Marc Per Declarar I Descriure mètriques De Qualitat De La Traducció [MQM: A framework for declaring and describing translation quality metrics]. *Tradumàtica*, *12*, 455-463. https://doi.org/10.5565/rev/tradumatica.77

Lu, B. H., Lin, Y. H., Lee, E. S. A., & Tsai, R. T. H. (2024). Enhancing Taiwanese Hokkien dual translation by exploring and standardizing of four writing systems. In *proceedings of the LREC-COLING 2024* (pp. 6077-6090). https://aclanthology.org/2024.lrec-main.538/

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabenamualu, S. K., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O. E., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., & ... Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2144-2160). https://doi.org/10.18653/v1/2020.findings-emnlp.195

Nida, E. A. (1964). *Toward a science of translating: With special reference to*

principles and procedures involved in Bible translating. Brill Archive.

OpenAI. (2025). ChatGPT (o4-mini) [Large language model]. https://chat.openai.com/chat

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Peng, C. Y. (2021). *Mediatized Taiwanese Mandarin: Popular culture, masculinity, and social perceptions*. Springer Singapore. https://doi.org/10.1007/978-981-15-4222-0

Pym, A. (2023). *Exploring translation theories*. Routledge.

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing* (EMNLP) (pp. 4512-4525). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, *44*(3), 393-401. https://doi.org/10.1162/COLI_a_00322

Sandel, T. L. (2002). Kinship address: Socializing young children in Taiwan. *Western Journal of Communication*, *66*(3), 257-280. https://psycnet.apa.org/record/2002-15743-005

Schreier, M. (2012). *Qualitative content analysis in practice*. Sage. https://doi.org/10.4135/9781529682571

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv*. https://doi.org/10.48550/arXiv.1511.06709

Sharma, D., & Deo, A. (2009). Contact-based aspectual restructuring: A critique of the aspect hypothesis. *Queen Mary's Occasional Papers Advancing Linguistics*, *12*, 1-29. https://www.qmul.ac.uk/arts/media/sllf-new/department-of-linguistics/12-QMOPAL-Sharma-Deo.pdf

Somers, H. (2005). Round-trip translation: What is it good for? In T. Baldwin, J. Curran, & M. van Zaanen (Eds.), *Proceedings of the Australasian language technology workshop* (pp. 127-133). ALTA.

Toury, G. (2012). *Descriptive translation studies and beyond*. John Benjamins.

Venuti, L. (2017). *The translator's invisibility: A history of translation* (3rd ed.). Routledge.

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*(1), 98-118. https://doi.org/10.1093/llc/fqt031

Walkowiak, T., & Gniewkowski, M. (2019). Evaluation of vector embedding models in clustering of text documents. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the international conference on recent advances in natural language processing (RANLP 2019)* (pp. 1304-1311). INCOMA. https://doi.org/10.26615/978-954-452-056-4_149

Wang, H., & Wang, X. (2025). Enhancing translation accuracy and learning outcomes for low-resource languages: AI fine-tuning and learner adoption factors. *System*, *134*, 103807. https://doi.org/10.1016/j.system.2025.103807

Zeitoun, E., & Wu, C. H. (2006). An overview of reduplication in Formosan languages. In Y. L. Chang, M. J. Huang, & D. A. He (Eds.), *Streams converging into an ocean: Festschrift in honor of Professor Paul Jen-kuei Li on his 70th birthday* (pp. 97-142). Institute of Linguistics, Academia Sinica.

Zhang, J. (2014). Tones, tonal phonology, and tone sandhi. In C. T. J. Huang, Y. H. A. Li, & A. Simpson (Eds.), *The handbook of Chinese linguistics* (pp. 443-

464). Wiley-Blackwell. https://doi.org/10.1002/9781118584552.ch17

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. *arXiv*. https://doi.org/10.48550/arXiv.1904.09675

Zhao, W., Wang, L., Shen, K., Jia, R., & Liu, J. (2019). Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv*. https://doi.org/10.48550/arXiv.1903.00138

Appendix A

Translation Prompt

You are a professional translator with expertise in Taiwanese (Tâi-gí) and Mandarin Chinese.

Task: Translate the following Taiwanese text into natural, fluent Mandarin Chinese.

Important guidelines:

- Maintain the original meaning as closely as possible
- Preserve cultural nuances where appropriate
- Use standard Mandarin Chinese orthography
- Maintain the original text's tone and register

Taiwanese text to translate:

[INPUT_TEXT]

Appendix B

Back-Translation Prompt

You are a professional translator with expertise in Taiwanese (Tâi-gí) and Mandarin Chinese.

Task: Translate the following Mandarin Chinese text into natural, fluent Taiwanese (Tâi-gí).

Important guidelines:

- Maintain the original meaning as closely as possible
- Use Taiwanese orthography with Han characters where appropriate
- Preserve the tone and register of the original text
- When multiple translations are possible, choose the most natural Taiwanese expression

Mandarin text to translate:

[INPUT_TEXT]