Ara-HOPE: Human-Centric Post-Editing Evaluation for Dialectal Arabic to Modern Standard Arabic Translation

Abdullah Alabdullah[*1], **Lifeng Han**[2], **Chenghua Lin**[3],

[1]School of Informatics, University of Edinburgh, The United Kingdom
[2]Leids Universitair Medisch Centrum & LIACS, Universiteit Leiden, NL
[3]School of Computer Science, University of Manchester, The United Kingdom

Correspondence: a.alabdullah@sms.ed.ac.uk & l.han@lumc.nl,liacs.leidenuniv.nl

Abstract

Dialectal Arabic to Modern Standard Arabic (DA-MSA) translation is a challenging task in Machine Translation (MT) due to significant lexical, syntactic, and semantic divergences between Arabic dialects and MSA. Existing automatic evaluation metrics and general-purpose human evaluation frameworks struggle to capture **dialect-specific MT errors**, hindering progress in translation assessment. This paper introduces *Ara-HOPE*, a human-centric post-editing evaluation framework designed to systematically address these challenges. The framework includes a five-category error taxonomy and a decision-tree annotation protocol. Through comparative evaluation of three MT systems (Arabic-centric Jais, general-purpose GPT-3.5, and baseline NLLB-200), Ara-HOPE effectively highlights systematic performance differences between these systems. Our results show that dialect-specific terminology and semantic preservation remain the most persistent challenges in DA-MSA translation. Ara-HOPE establishes a new framework for evaluating Dialectal Arabic MT quality and provides actionable guidance for improving dialect-aware MT systems. For reproducibility, we make the annotation files and related materials publicly available at `https://github.com/abdullahalabdullah/Ara-HOPE`

1 Introduction

This paper enhances Dialectal Arabic to Modern Standard Arabic (DA-MSA) translation quality assessment through a human-centric evaluation framework, addressing key gaps in current methods for under-resourced language pairs.

Dialectal Arabic (DA) refers to the informal varieties of Arabic used in everyday communication, which vary significantly across regions. In contrast, Modern Standard Arabic (MSA) is the formal variety used in writing, education, and traditional media (Diab et al., 2010).

Unlike interlingual translation tasks (e.g. translating from English to Arabic), DA-to-MSA translation is a dialect normalization task that introduces additional challenges arising from divergences between Arabic dialects and MSA. Key differences between DA and MSA that affect MT include: (i) **Orthographic differences**: Dialects do not follow standardized spelling rules and the same word may appear in different forms, making automatic text normalization difficult for NLP systems (Alhafni et al., 2024). (ii) **Morphological differences**: Morphology is how words change form to express features such as tense or gender. While MSA has a rich, complex, and standardized morphology, spoken dialects often simplify these systems by omitting rules or using reduced forms (Kirchhoff et al., 2006). (iii) **Lexical differences**: Dialectal vocabulary often includes slang and idiomatic expressions that are typically absent in MSA (Hadj Mohamed et al., 2023). (iv) **Syntactic differences**: Syntax, sentence structure, and word order in many dialects differ from MSA (Biadsy et al., 2009). (v) **Code-switch**: Speakers often mix DA and MSA, and sometimes even foreign words, within a single sentence. This adds more complexity when trying to build systems that automatically translate from DA to MSA (Hamed et al., 2025). These challenges make DA-MSA translation a complex task requiring specialized approaches grounded in a sound understanding of common error types in DA-MSA translation systems.

In this paper we introduce a post-editing human evaluation framework that offers the granularity needed to identify systemic weaknesses in DA-MSA machine translation systems. Unlike general-purpose frameworks, our proposed framework (Ara-HOPE) targets translation errors that result from DA-specific translation challenges.

[*] The previous affiliation for Abdullah Alabdullah (where most of this research was completed) is the University of Manchester, Manchester, M13 9PL, The United Kingdom.

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 157–171
March 29, 2026. ©2026 Association for Computational Linguistics

2 Related Work

2.1 Advancements in Dialectal Arabic Translation

Neural architectures have transformed DA-MSA translation by capturing contextual dependencies and complex syntactic and semantic relationships, effectively modeling dialectal variations and producing more fluent and accurate translations (Baniata et al., 2018). The emergence of large parallel corpora has further advanced Neural Machine Translation by mitigating data scarcity and enabling training on dialectally diverse data. A key example is the MADAR corpus, which contains parallel translations from 25 Arabic city dialects (Bouamor et al., 2018). Recent studies have compared neural architectures, including encoder–decoder models like NLLB-200 and decoder-only models like GPT-3 and GPT-4o (Team et al., 2022; Brown et al., 2020; Alabdullah et al., 2025). Decoder-only models have demonstrated better performance at preserving cultural context (Yakhni and Chehab, 2025).

The emergence of LLMs trained on large and diverse multilingual data, and further optimized through instruction tuning, has enabled models to follow natural-language instructions and generate appropriate outputs for a wide range of tasks, including machine translation, without costly task-specific fine-tuning (Brown et al., 2020). This paradigm is known as In-Context Learning (ICL), where the desired task is specified directly in the prompt, and the model infers the mapping from the provided context. In zero-shot ICL, the model is instructed to translate from a source to a target language without providing any in-prompt examples. Zero-shot prompting has been particularly effective when parallel data is scarce and is widely used in DA-MSA shared tasks such as OSACT (Atwany et al., 2024) and NADI (Abdul-Mageed et al., 2024).

While multilingual LLMs capture general linguistic features through large-scale multilingual pretraining, new Arabic-specialized LLMs like Jais improved the handling of dialectal nuances and cultural references (Sengupta et al., 2023; Mousi et al., 2025), producing more natural and contextually appropriate translations. Despite these advances, neural models continue to underperform on DA translation due to persistent challenges with dialectical nuances and culturally embedded expressions (Mousi et al., 2025; Alabdullah et al., 2025).

2.2 Dialectal Arabic Translation Evaluation

Traditional automatic evaluation metrics such as BLEU (Post, 2018) and METEOR (Banerjee and Lavie, 2005) are limited for DA-MSA translation, as they rely on lexical overlap and perform poorly on morphologically rich languages like Arabic. Bouamor et al. (2014) proposed AL-BLEU, an extension of BLEU that assigns partial credit for stem and morphological matches, yielding better correlation with human judgment than standard metrics. However, AL-BLEU remains a lexical-overlap metric and fails to capture semantic adequacy, particularly an issue for syntactically flexible languages like Arabic. While metrics like BLEU allow evaluation against multiple reference translations, these are challenging to produce. As a result, these metrics favor literal translations with high lexical overlap over contextually appropriate ones that better reflect human judgment, but differ lexically from the reference.

In low-resource settings such as dialectal Arabic, neural evaluation metrics like BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020) also face challenges. Falcão et al. (2024) showed that COMET's performance on under-resourced languages is constrained by imbalanced training data. These metrics rely on pretrained models which typically lack sufficient dialectal Arabic training data, leading to lower-quality embeddings.

Human evaluation frameworks such as the Multidimensional Quality Metrics (MQM) assess fine-grained translation errors like omissions and register mismatches (Lommel et al., 2013, 2024). While MQM allows for more precise diagnostic error analysis due to its detailed error taxonomy, this comes at the cost of increased complexity in the annotation framework, requiring extensive annotator training, which can be costly and difficult to achieve for low-resource languages (Kocmi et al., 2024). Moreover, DA translation requires a specialized human evaluation framework that captures the most impactful error types and minimizes subjectivity in assessing translation quality for this specific task, with only minimally trained annotators. To further advance research in this direction, we build our methodology on the HOPE metric (Gladkoff and Han, 2022). HOPE is a task-oriented evaluation framework designed to address the limitations of both automatic evaluation metrics (e.g. BLEU) and highly fine-grained but complex human evaluation frameworks such as MQM. Ara-HOPE is human-centric because it only incorporates eight evaluation criteria that capture the most critical and recurring errors in translation between the Levantine Arabic dialect and MSA, reflecting translation quality as perceived by native speakers. The assigned error scores correspond to the post-editing effort required to bring the translation to an acceptable quality.

3 Methodology Design

3.1 Framework Development

Ara-HOPE Error Taxonomy Design: Developing a robust human evaluation framework requires clearly defined objectives. For DA-MSA translation, this requires an error taxonomy tailored to the specific challenges of this language pair. While the HOPE framework (Gladkoff and Han, 2022) offers a foundation for general post-editing translation assessment, we adapt the general error types in HOPE to capture the specific challenges of DA-MSA translation. To minimize subjective judgment between annotators, we also reduce the severity scoring range to three levels: from 0 (no errors) to 2 (major error). Our proposed taxonomy is designed to evaluate translation quality, identify system weaknesses, and guide system improvements, while remaining usable by native DA speakers without requiring extensive annotation training. To achieve this, we employ a direct quality estimation approach that evaluates predefined aspects of translation quality (e.g. fluency) at the segment level using discrete severity scores. This design makes our framework easier to train annotators on and faster to apply than MQM.

Design Principles: To ensure a theoretically sound and practical taxonomy, we developed guidelines grounded in best practices in translation quality assessment (Han, 2020; Rivera-Trigueros, 2022) and refined them through iterative feedback from focus groups with native DA speakers. The finalized guidelines comprise six core principles: (1) **Identifiability:** Errors must be detectable by minimally trained native dialect speakers to ensure consistent evaluation. (2) **Distinguishability:** Categories should have minimal conceptual overlap, with clear definitions to avoid confusion. (3) **Actionability:** Error classification must enable targeted system improvements through quantifiable error severity and clear mapping to issues in the translation system. (4) **Comprehensiveness:** The taxonomy must capture all major errors common in DA-MSA MT, including dialect-specific challenges. (5) **Relevance:** Categories must address issues unique to DA-MSA translation rather than pure general translation errors. (6) **Usability:** The taxonomy should remain manageable in size, balancing high-level categories with sufficient granularity.

Taxonomy Structure: The Ara-HOPE taxonomy categorizes DA-MSA translation errors into three hierarchically structured classes: (1) **Fluency Error (FLU):** Grammatical or linguistic errors in the MSA translation, independent of the DA source sentence. (2) **Meaning Transfer Error:** Failures to accurately preserve source meaning. This category includes Proper Name (PRN) errors, referring to incorrect translations of names of people, places, or organizations; Dialect-Specific Term (TRM) errors, which involve untranslated or mistranslated dialectal expressions that alter meaning; and General Semantic Mistranslation (GSMIS) errors, which includes omissions, additions, or other semantic changes. (3) **Adaptation Error (ADP):** Translations that are unnatural or contextually inappropriate in tone, style, or intent.

GSMIS captures global meaning changes arising from the model's inability to handle DA's contextual dependencies and includes all meaning changes not covered by PRN or TRM. This category was introduced after observing that LLMs often produce semantic distortions beyond proper name or dialect-specific term mistranslations.

Decision Tree Implementation While a list format effectively outlined the taxonomy, it was later reformatted as a decision tree to reduce cognitive load for our minimally trained evaluators. The decision tree guides annotators through hierarchical error categories, as shown in Figure 5 in the Appendix. An English translated version of the tree is provided in Figure 6 in the Appendix.

This structure simplifies annotation by providing step-by-step guidance, helping evaluators understand the evaluation process, handle multiple error types systematically, and make consistent judgment at each level. The tree begins with the three primary categories (Fluency, Meaning Transfer, and Adaptation) and expands into more granular subcategories. Each node presents a yes/no question to minimize subjective judgments. Since some error types (e.g. Dialect-Specific Term) can be challenging to distinguish from other categories, we provided annotators with a table of annotation guidelines with illustrative examples to clarify the distinctions between error types. The guidelines can be found in Figure 7 in the Appendix.

Practical initial testing demonstrated Ara-HOPE's effectiveness in addressing dialect-specific translation challenges and improving annotation consistency. The decision-tree structure also proved especially effective in guiding evaluators and enhancing usability.

3.2 Dataset Preparation

Our human evaluation experiment utilized 200 tweets from the Levantine development set of the Dial2MSA-Verified dataset (Khered et al., 2025), a high-quality parallel corpus for DA-MSA translation in the social media domain. It extends the original Dial2MSA dataset (Mubarak, 2018) and applies automated corrections and human evaluation by native speakers to produce reliable MSA references.

This tweets dataset was chosen for its comprehensive representation of DA-MSA translation challenges, including: **Lexical variations:** colloquial and multi-word expressions. **Semantic shifts:** cultural references and context-dependent meaning. **Orthographic variations:** common in Levantine social media text.

The tweets in this dataset span a wide range of tones, from casual tweets to heated discourse, ensuring thorough testing of MT systems' ability to handle DA translation, while preserving sentiment. Moreover, the dataset allows for testing Ara-HOPE's capacity to capture nuanced dialect-specific translation errors.

Our 200-example subset size exceeds the lower bound below which translation quality estimates become unreliable, addressing concerns that a very small sample may lead to unreliable error analysis (Gladkoff et al., 2022).

3.3 Translation Systems Selection

For this human evaluation experiment, we selected three translation systems using three predefined criteria, which ensure a robust comparative analysis of DA-MSA translation performance: (i) **Reliability and Proven Performance**: Each system has demonstrated strong performance on DA-MSA translation in prior work. (ii) **Architectural Diversity**: The systems represent distinct model families (encoder-decoder vs. decoder-only) and differ in pretraining data sources, encouraging diverse outputs and reducing redundancy. (iii) **Performance Scope**: We include both advanced and baseline systems to contrast general-purpose and Arabic-specialized approaches.

The Selected Systems Are: **Jais**: A state-of-the-art Arabic-centric model trained with approximately one-third of its training tokens drawn from Arabic data, excelling at capturing DA-MSA nuances (Sengupta et al., 2023). **GPT-3.5**: A general-purpose multilingual LLM. We will use this model to benchmark multilingual systems against an Arabic-specialized model. This model was introduced as an improvement over its predecessor, GPT-3 (Brown et al., 2020). **NLLB-200 3.3B**: A multilingual baseline MT system, providing contrast and revealing errors that more advanced systems may avoid (Team et al., 2022).

Each of the three systems was prompted to translate the same 200 examples. A zero-shot instruction setup was used to fairly evaluate each model's baseline capabilities without fine-tuning or advanced prompting configuration. Jais and GPT-3.5 received consistent prompts (in Arabic for Jais and English for GPT-3.5) while NLLB-200 required no prompting due to its built-in MT task support.

4 Implementation

4.1 Annotator Selection and Training

Two native speakers of Syrian Levantine Arabic with advanced proficiency in MSA volunteered as annotators. Both hold undergraduate degrees in Arabic Language, enabling them to identify violations of MSA grammatical rules. Recruiting two annotators allowed for inter-annotator agreement (IAA) analysis while keeping the annotators group size manageable.

Annotators received initial training through a 25-minute pre-recorded video and supplementary materials. Training covered: (1) an **Error Taxonomy Guide**: providing detailed definitions of error types, alongside diverse examples of correct and incorrect DA-MSA translations, and (2) a **Decision Tree Workflow**: annotators first read the source and gold translation to establish the intended meaning, then highlighted errors by severity starting with the most critical, and finally used the decision tree (Figure 6 in the Appendix) to classify errors systematically.

Offline support was available throughout the annotation period. Each annotator spent approximately 12 hours on their tasks, supported by multiple feedback sessions to ensure clarity and consistency.

4.2 Pilot Testing and Framework Refinement

Before full-scale evaluation, a pilot study was conducted with one annotator annotating 10 DA-MSA examples. The primary goal was to test the framework's usability and effectiveness. The feedback received focused on: (1) **Instruction Clarity**: The annotator initially struggled to distinguish between meaning transfer and adaptation errors. The definitions were revised accordingly. (2) **Questionnaire Usability**: The layout was changed to ensure ease of use and reduce cognitive load. (3) **Severity Scale Adjustments**: Confusing mid-range scores were removed, leading to refinement of the scoring scale.

The pilot test confirmed the framework's overall effectiveness and identified minor improvements needed to enhance clarity and usability. Revisions ensured annotators could complete evaluation efficiently without sacrificing accuracy.

4.3 Questionnaire Design

A structured Excel sheet was created to support annotation and data analysis. Each sheet contained three sets of columns, one per translation engine. Each annotator evaluated 600 translations in total (200 for each translation engine). Figure 1 shows two annotated examples for the Jais system. The first three columns show the DA sentence, the MSA gold translation, and the proposed machine translation. Columns 4-8 represent the five Ara-HOPE error categories, with annotators assigning severity scores from 0 (no error) to 2 (major error). Empty cells indicate a score of 0, and the final column sums the error scores for each sentence.

Annotators assessed adaptation errors only when no meaning transfer errors (PRN, TRM, GSMIS) were present, as style or tone evaluation is irrelevant when meaning is lost. To prevent over-penalizing systems that preserved meaning but erred in adaptation, adaptation scores were weighted at 50% in the segment total error score (SEGS) calculation.

Three physical copies of the Excel sheet were printed (one for each system) as annotators preferred working on paper. Each contained the same DA-MSA pairs, differing only in the translation engine output. Annotators marked scores manually, and data were later transferred to the digital Excel template.

5 Results Analysis

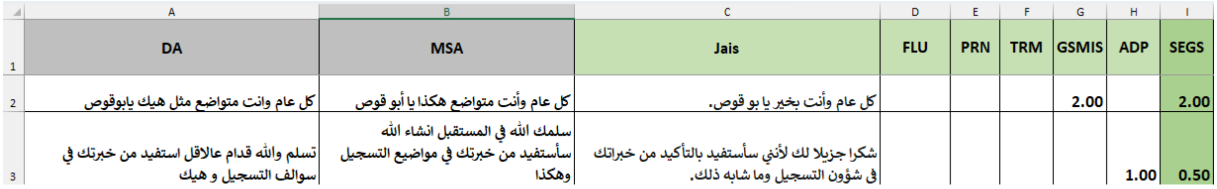
5.1 Inter Annotator Agreement

Inter-Annotator Agreement (IAA) measures consistency between evaluators, with high scores indicating a reliable and reproducible annotation framework. In our study, we employed Quadratic Weighted Cohen's Kappa (QWK) because it effectively handles ordinal severity ratings by assigning greater penalties to larger disagreements, whereas, the ordinary Cohen's κ treats all disagreements equally, unweighted κ, e.g. 0 vs 1 = 0 vs 2. In addition, quadratic weights are standard for ordinal ratings (Cohen, 1968). This makes it particularly suitable for error severity judgment on our 0-2 scale, where the magnitude of disagreement matters. The IAA scores are presented in Table 1.

As shown in Figure 6, Meaning Transfer comprises three fine-grained error types (PRN, TRM, and GSMIS). Accordingly, for Meaning Transfer IAA, we compute inter-annotator agreement on an aggregated severity score obtained by summing PRN, TRM, and GSMIS. This reflects annotator agreement on overall meaning-transfer impact rather than on individual sub-error categories. Similarly, SEGS represents the segment-level total error score (produced by summing all five error types' scores), and the corresponding IAA reflects agreement on overall segment-level error severity.

The IAA scores varied meaningfully across error categories: Fluency showed strong agreement for most MT systems, likely due annotators' Arabic language expertise, which minimized inaccurate judgments. Meaning Transfer achieved robust agreement, likely due to the clear criteria regarding preservation of original meaning, as established in the Ara-HOPE taxonomy decision tree (Figure 6). Those definitions were particularly refined before the full evaluation during our pilot testing and feedback sessions. This consistency indicates a shared understanding of meaning preservation requirements among our annotators. Adaptation yielded the lowest agreement (0.122-0.28) due to its inherently subjective nature.

A translation has an adaptation error when it seems unnatural or inappropriate to the context, situation, or intent of the source. Unlike fluency or meaning-transfer errors, which can often be localized to specific parts of the translation, adaptation errors require more holistic, and usually subjective judgment. One example is the dialectal sentence:



	A	B	C	D	E	F	G	H	I
1	DA	MSA	Jais	FLU	PRN	TRM	GSMIS	ADP	SEGS
2	كل عام وانت متواضع مثل هيك يابوقوص	كل عام وأنت متواضع هكذا يا أبو قوص	كل عام وأنت بخير يا بو قوص.				2.00		2.00
3	تسلم والله قدام عالاقل استفيد من خبرتك في سوالف التسجيل و هيك	سلمك الله في المستقبل انشاء الله سأستفيد من خبرتك في مواضيع التسجيل وهكذا	شكرا جزيلا لك لأنني سأستفيد بالتأكيد من خبراتك في شؤون التسجيل وما شابه ذلك.					1.00	0.50

Figure 1: Example layout of the annotation questionnaire used for evaluating DA-MSA translations across fluency, meaning transfer, and adaptation error categories. To avoid unfairly penalizing systems that preserved meaning but made adaptation errors, adaptation scores were weighted half in the segment total error score calculation.

عم بحكي من واقع تجربة كانت رح تقضي ع حلم حياتي بس الحمد لله الحب اقوى ("I am speaking from experience. It almost destroyed my life's dream, but thank God, love is stronger"), which the Jais model translated as

أنا أتحدث بناءً على تجربتي الشخصية التي كادت تدمر حلمي الأكبر، ولكن بحمد الله، أثبت الحب أنه أقوى بكثير. ("I speak from my personal experience, which almost destroyed my biggest dream, but thank God, love proved to be much stronger.")

While both annotators agreed that the translation is fluent and preserves the source meaning, one judged it to be slightly less natural in MSA (ADP = 1, minor adaptation error), whereas the other annotator found it acceptable and contextually appropriate.

The annotator disagreement likely stems from the rendering of phrases suchas بس الحمد لله الحب اقوى ("But thank God, love is stronger") which was translated as ولكن بحمد الله، أثبت الحب أنه أقوى بكثير. ("But thank God, love proved to be much stronger"), which may be seen as stylistically over-emphatic or slightly less natural in MSA for this context.

Error Type	Jias	GPT3.5	NLLB200
Fluency	0.507	0.552	0.368
Meaning Transfer	0.529	0.629	0.554
Adaptation	0.171	0.122	0.280
SEGS	0.608	0.629	0.500

Table 1: Quadratic Weighted Kappa (QWK) scores across models and error types. Meaning Transfer IAA is computed on aggregated PRN+TRM+GSMIS severity; SEGS reflects total segment-level error severity.

The IAA scores for the three systems in our study on the segment-level total error score (SEGS) was 0.5 to 0.629, indicating reasonably consistent human evaluation.

Prior work, including (Landis and Koch, 1977), considers kappa scores in the range of 0.41-0.60 to indicate a moderate level of agreement and 0.61-0.80 to indicate substantial agreement. Nevertheless, it is important to note that the interpretation of standard Kappa and Quadratic Weighted Kappa varies considerably.

Aggregating fine-grained error categories into a composite severity score reduces sparsity and stabilizes the marginal distributions, which is known to improve reliability of composite measures (Fleiss et al., 2003). Moreover, Quadratic Weighted Kappa assigns smaller penalties to near disagreements (Cohen, 1968), so cross-category disagreements often translate into small ordinal differences after aggregation. Consequently, agreement computed on aggregated segment-level scores can exceed that of individual categories.

5.2 Quantitative Error Analysis

The quantitative analysis below presents a comparison of MT system performance, highlighting differences across key evaluation criteria. This assessment identifies each system's strengths and weaknesses, as well as both unique and shared challenges across systems.

Error Severity Analysis This analysis examines translation error severity for 205 sentences produced by the three MT systems using the accumulated sentence total error scores (SEGS). It is essential for understanding how each model handles DA-MSA translation and for identifying systems that generate higher-quality outputs that require minimal post-editing.

SEGS is calculated as the sum of the scores assigned to each of the five error types per sentence, yielding a range of 0-4 per sentence. For comparison, SEGS values were grouped as follows: segments requiring no editing (SEGS = 0), segments with minor errors ($0.5 < \text{SEGS} \leq 1$), and segments with major errors ($\text{SEGS} > 1$).

Figure 2 shows clear differences among the systems. For Jais, 36% of segments required no editing, 34% had minor errors, and 30% had major er-

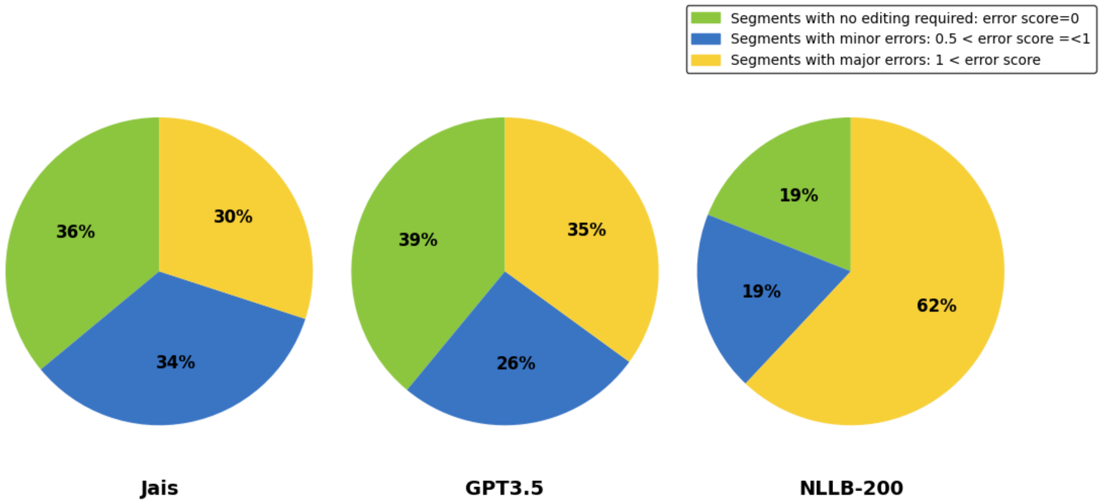


Figure 2: Comparison of error severity distribution among Jais, GPT-3.5, and NLLB-200, highlighting proportions of major, minor, and no-error translations.

rors. GPT-3.5 performed slightly better, with 39% of segments requiring no editing, fewer minor errors (26%), but slightly more major errors (35%). NLLB-200 performed the worst, with only 19% of segments requiring no editing and 62% containing major errors.

These results reflect differences in model architecture and training. Jais, as an Arabic-centric model, handles DA-specific nuances well but still struggles with complex segments. GPT-3.5's general-purpose design provides balanced performance but lacks Arabic specialization. The strong performance of Jais and GPT-3.5 underscores the relative strength of decoder-only models over encoder-decoder models like NLLB-200.

Error Pattern Examining the error distributions in Figure 3 provides key insights into each system's approach to DA-MSA translation challenges. NLLB-200 exhibits a high frequency of dialect-specific term (TRM) and semantic preservation (GSMIS) errors, highlighting its encoder–decoder architecture's difficulty with idiomatic expressions, cultural references, and context-dependent meanings specific to Levantine Arabic. These issues stem from limited domain-specific tuning and insufficient exposure to regional culturally embedded expressions, which are essential for accurate meaning transfer.

Jais, despite its Arabic-specialized pretraining, struggles with proper names (PRN). This might be because the system tends to over-normalize or excessively standardize names transliterated in social media DA content, leading to the loss of original or intended forms. However, its strong performance with dialect-specific terminology reflects effective handling of DA morphological variations, likely due to its extensive pretraining on Arabic data.

GPT-3.5 shows a balanced error distribution across many categories, suggesting that its large multilingual pretraining and sheer parameter size help offset the lack of explicit Arabic dialectal training. This is particularly evident in challenging cases involving code-mixing (where words from two languages or dialects are used in the same sentence) and pragmatic shifts, which require interpreting meaning based on context, tone, or speaker intent rather than literal words. These cases require advanced comprehension, and the model performs well in them even though it was not specifically trained on Arabic dialects.

The consistently low fluency (FLU) error rates across systems indicate that syntactic reconstruction from DA to MSA is less challenging than lexical-semantic transfer[1]. Low adaptation (ADP) error rates across all systems are largely due to this error type being assessed only when meaning transfer errors (PRN, TRM, GSMIS) are absent. When a text fails to convey its intended meaning, evaluating its stylistic or cultural appropriateness becomes less relevant.

A complementary view is provided in Figure 4, which shows the exact error distribution for each system. The total error scores for all sentences

[1] Lexical-semantic transfer refers to the mapping of words and their intended meaning from one dialect or language variety to another.

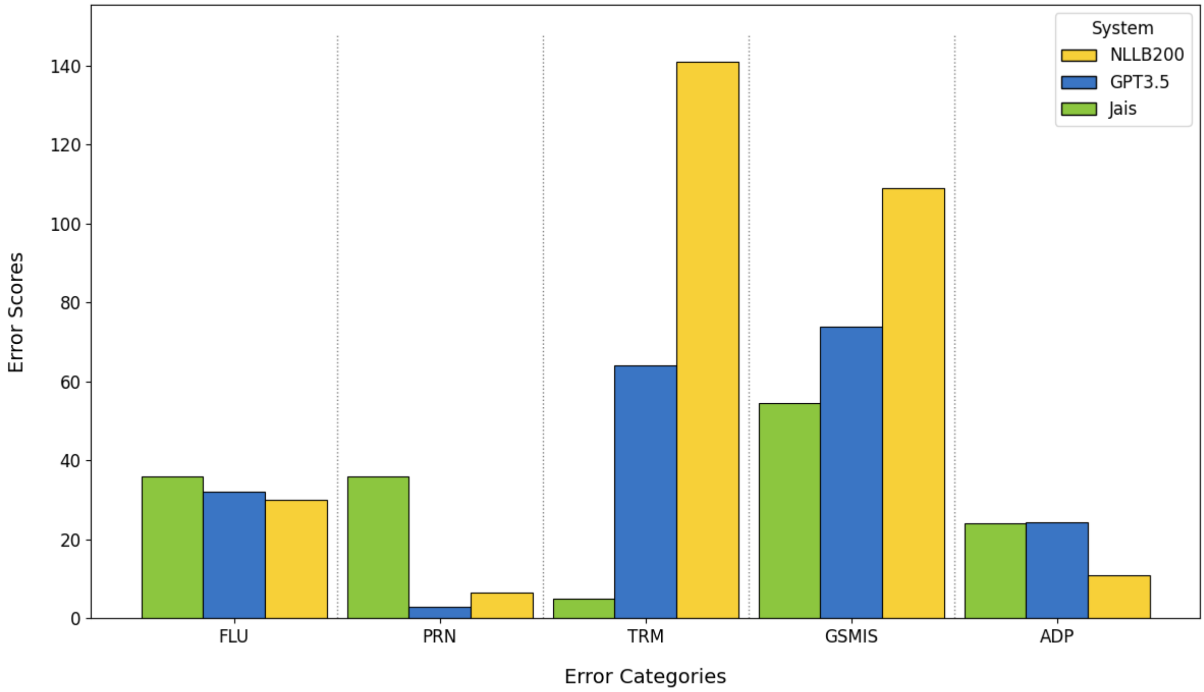


Figure 3: Comparison of Jais, GPT-3.5, and NLLB-200 models' error scores across error categories

(SEGS). It is clear that NLLB-200 has a significantly higher total error score (297.50) compared to GPT-3.5 (196.25) and Jais (187.50), with Jais producing the least meaning transfer errors. Notably, TRM and GSMIS errors constitute the majority of total errors across all systems, indicating that future MT development for DA-MSA should prioritize improvements in semantic accuracy and dialect-specific terminology handling.

5.3 Qualitative Error Analysis

This qualitative analysis offers concrete examples of two common error types to illustrate how translation quality varies across MT systems, focusing on challenging linguistic phenomena.

5.3.1 Analysis of Dialectal Terminology Errors

Addressing dialect-specific terms (TRM) is a core challenge in DA-MSA translation due to vocabulary mismatches between informal dialects and formal Arabic. Table 2 presents a sentence containing a unique dialect-specific term. The word يجنن (”amazing” in Syrian Arabic) lacks a direct MSA equivalent, requiring systems to interpret its meaning contextually instead of relying on word-for-word mapping. Unlike NLLB-200, Jais, with its Arabic-centric pretraining, was better at recognizing this regional idioms. GPT-3.5, though not DA-specialized, leverages its broad language understanding to connect dialect terms to contextually appropriate MSA words (translating يجنن as ممتاز ”excellent”). These findings demonstrate that success with TRM errors depends heavily on exposure to diverse, dialect-rich data that treats dialectal phrases as meaningful units to be interpreted contextually, rather than treating them as isolated words.

Source DA: اليوم الجو يجنن، لازم نطلع نتمشى **Gold MSA:** اليوم الطقس رائع، يجب أن نخرج للنزهة “The weather is wonderful today; we should go out for a walk.”
Jais: اليوم الطقس رائع، يجب أن نذهب للنزهة “The weather is wonderful today; we should go for a walk.”
GPT-3.5: اليوم الطقس ممتاز، يجب أن نخرج للنزهة “The weather is excellent today; we should go out for a walk.”
NLLB-200: اليوم الطقس مجنون، يجب أن نخرج “Today’s weather is crazy; we should go out for a walk.”

Table 2: Evaluation of how models handle dialect-specific terms (TRM), highlighting translation challenges in translating informal expressions to MSA.

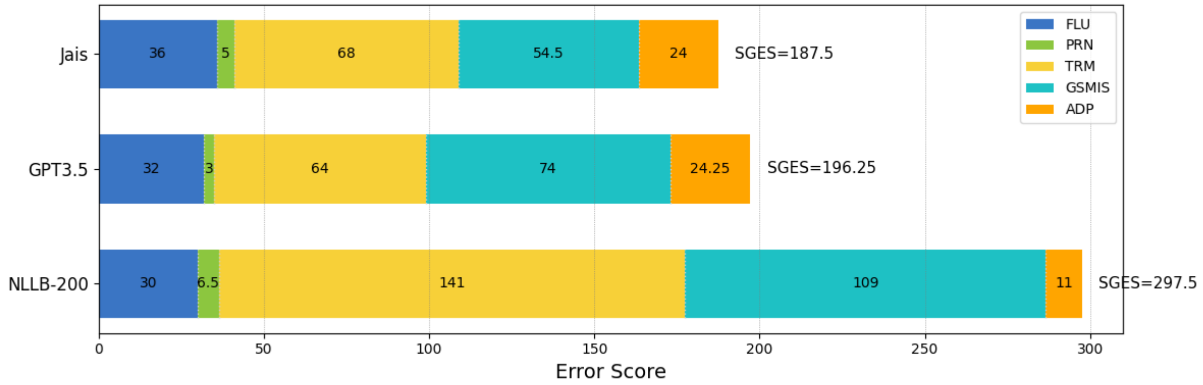


Figure 4: Visualization of accumulated error scores across fluency, meaning transfer, and adaptation error categories for Jais, GPT-3.5, and NLLB-200 in DA-MSA translation.

Source DA: شو قصتك ليش معصب **Gold MSA:** ما الأمر؟ لماذا تبدو غاضباً؟ "What's wrong? Why do you look angry?"
Jais: ما الخطب؟ لماذا تبدو غاضباً "What's the matter? Why do you look angry?"
GPT-3.5: ماذا حدث؟ لماذا تبدو غاضباً؟ "What happened? Why do you look angry?"
NLLB-200: ما هي قصتك؟ لماذا أنت غاضب؟ "What is your story? Why are you angry?"

Table 3: Comparison of model performance on Adaptation (ADP) errors.

5.3.2 Analysis of Adaptation Errors

Table 3 presents an example illustrating how the different systems preserved the intent and tone of the source when translating to MSA. ADP errors highlight the gap between DA's context-dependent expressions and MSA's formality. For instance, translating شو قصتك (casual "What's wrong?") requires capturing the speaker's intent rather than just the literal words. NLLB-200 translates this as ما هي قصتك ("What is your story?"), focusing on a literal interpretation. Jais, benefiting from exposure to both dialect and MSA data, uses the conventional MSA phrase ما الخطب ("What's the matter?"). GPT-3.5 chooses the phrasing ماذا حدث ("What happened?"). These differences show that reducing ADP errors requires systems to prioritize intent and context over lexical mapping, treating DA as a distinct communication style with its own contextual rules rather than simply a variation of MSA.

Overall, our analyses emphasize that effective DA-MSA translation depends on training strategies that prioritize (1) comprehension of dialect phrases and (2) preservation of speaker intent across registers. The qualitative analysis above shows that low-resource MT is a creative task, involving sub-tasks like sentiment analysis and formality adaptation, which go beyond simple lexical mapping.

6 Conclusion

This paper introduces the *Ara-HOPE* framework as a human-centric approach for evaluating DA-MSA translation, successfully fulfilling its intended objectives through a specialized error taxonomy, an efficient annotation workflow, and a comparative evaluation of different MT systems. The *five-category error classification* system effectively captures translation challenges unique to DA, while the *decision tree protocol* improves annotation consistency. Quantitative findings reveal significant differences in performance among Arabic-centric (Jais), general-purpose (GPT-3.5), and baseline (NLLB-200) systems, with dialect-specific terminology and semantic preservation identified as key challenges. By systematically addressing the complexities of DA-MSA translation assessment through rigorous human evaluation, Ara-HOPE establishes reproducible standards for Arabic MT assessment and provides actionable insights to guide future MT systems development [2].

[2] This work is in line with our DA-MSA MT work at (Alabdullah et al., 2025) where we examined LLM prompting vs finetuning for Levantine, Egyptian, and Gulf dialects to MSA translation.

7 Limitations

In this work, we only used zero-shot prompting to generate translations. Future research could explore human evaluation of translations produced using alternative prompting strategies, such as few-shot or chain-of-thought prompting. Additionally, the human annotation process was time-consuming. Future work could consider using LLM-as-a-judge approaches to partially or fully automate the annotation process. Our work focused on human evaluation, and we did not investigate the correlation between human judgment and automatic evaluation metrics, lexicon-based and neural-embedding based, like BLEU and BERTScore. We leave that for future work.

Acknowledgment

We would like to thank our volunteer human annotators, Heba Sado Jazea and Rania Mohammad Daher, for their invaluable time and effort throughout the annotation process. Their careful annotations and insightful feedback were essential to establishing our Ara-HOPE framework. We also thank the VarDial reviewers for the valuable comments.

References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2025. Advancing dialectal arabic to modern standard arabic machine translation. *Preprint*, arXiv:2507.20301.

Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor, and Nizar Habash. 2024. Exploiting dialect identification in automatic dialectal text normalization. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 42–54, Bangkok, Thailand. Association for Computational Linguistics.

Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. OSACT 2024 task 2: Arabic dialect to MSA translation. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 98–103, Torino, Italia. ELRA and ICCL.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Laith H. Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilises multitask learning (mtl). *Computational Intelligence and Neuroscience*, 2018. Publisher Copyright: © 2018 Laith H. Baniata et al.

Fadi Biadsy, Julia Hirschberg, and Nizar Habash. 2009. Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Semitic '09, page 53–61, USA. Association for Computational Linguistics.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for arabic mt evaluation. In *Conference on Empirical Methods in Natural Language Processing*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The madar arabic dialect corpus and lexicon. In *International Conference on Language Resources and Evaluation*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement. *Psychological Bulletin*, 70(4):213–220.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. Colaba: Arabic dialect annotation and processing. In *Lrec workshop on semitic language processing*, pages 66–74.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of* *the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical Methods for Rates and Proportions*. Wiley.

Serge Gladkoff and Lifeng Han. 2022. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.

Serge Gladkoff, Irina Sorokina, Lifeng Han, and Alexandra Alekseeva. 2022. Measuring uncertainty in translation quality evaluation (TQE). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1454–1461, Marseille, France. European Language Resources Association.

Najet Hadj Mohamed, Malak Rassem, Lifeng Han, and Goran Nenadic. 2023. AlphaMWE-Arabic: Arabic edition of multilingual parallel corpora with multiword expression annotations. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 448–457, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4561–4585, Abu Dhabi, UAE. Association for Computational Linguistics.

Chao Han. 2020. Translation quality assessment: a critical methodological review. *The Translator*, 26(3):257–273.

Abdullah Khered, Youcef Benkhedda, and Riza Batista-Navarro. 2025. Dial2MSA-verified: A multi-dialect Arabic social media dataset for neural machine translation to Modern Standard Arabic. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 50–62, Abu Dhabi, UAE. Association for Computational Linguistics.

Katrin Kirchhoff, Dimitra Vergyri, Jeff Bilmes, Kevin Duh, and Andreas Stolcke. 2006. Morphology-based language modeling for conversational arabic speech recognition. *Computer Speech & Language*, 20(4):589–608.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Hamdy Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. *OSACT*, 3:49.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

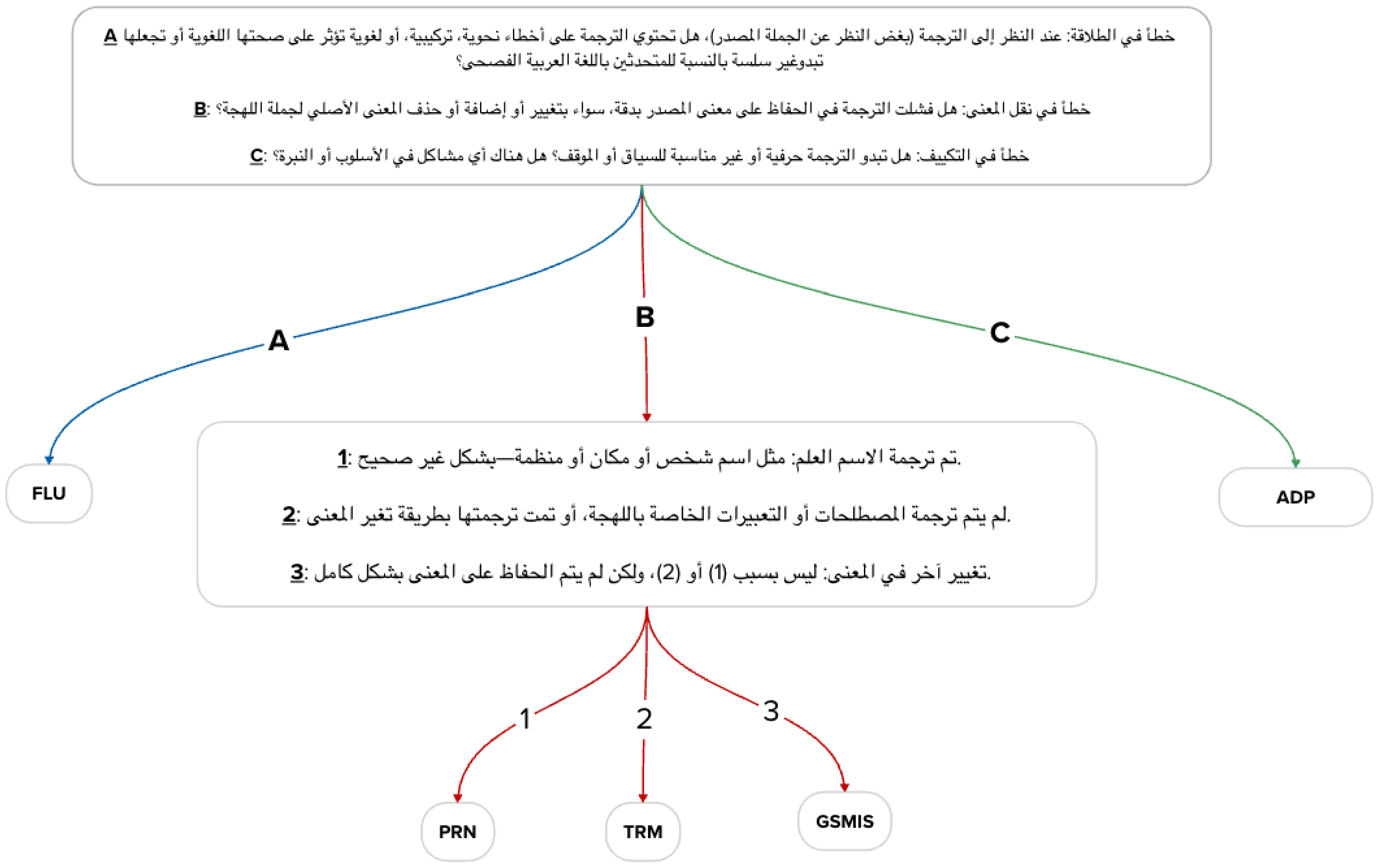
Irene Rivera-Trigueros. 2022. Machine translation systems and quality assessment: a systematic review. *Lang. Resour. Eval.*, 56(2):593–619.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Silvana Yakhni and Ali Chehab. 2025. Can LLMs translate cultural nuance in dialects? a case study on Lebanese Arabic. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 114–135, Abu Dhabi, UAE. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.



خطأ في الطلاقة: عند النظر إلى الترجمة (بغض النظر عن الجملة المصدر)، هل تحتوي الترجمة على أخطاء نحوية، تركيبية، أو لغوية تؤثر على صحتها اللغوية أو تجعلها **A**
تبدوغير سلسة بالنسبة للمتحدثين باللغة العربية الفصحى؟

خطأ في نقل المعنى: هل فشلت الترجمة في الحفاظ على معنى المصدر بدقة، سواء بتغيير أو إضافة أو حذف المعنى الأصلي لجملة اللهجة؟ :**B**

خطأ في التكييف: هل تبدو الترجمة حرفية أو غير مناسبة للسياق أو الموقف؟ هل هناك أي مشاكل في الأسلوب أو النبرة؟ :**C**

1: تم ترجمة الاسم العلم: مثل اسم شخص أو مكان أو منظمة—بشكل غير صحيح.

2: لم يتم ترجمة المصطلحات أو التعبيرات الخاصة باللهجة، أو تمت ترجمتها بطريقة تغير المعنى.

3: تغيير آخر في المعنى: ليس بسبب (1) أو (2)، ولكن لم يتم الحفاظ على المعنى بشكل كامل.

Figure 5: The Arabic version of the Ara-HOPE Annotation Decision Tree. A structured decision tree guiding annotators through error classification for evaluating DA-MSA translations using the Ara-HOPE framework.

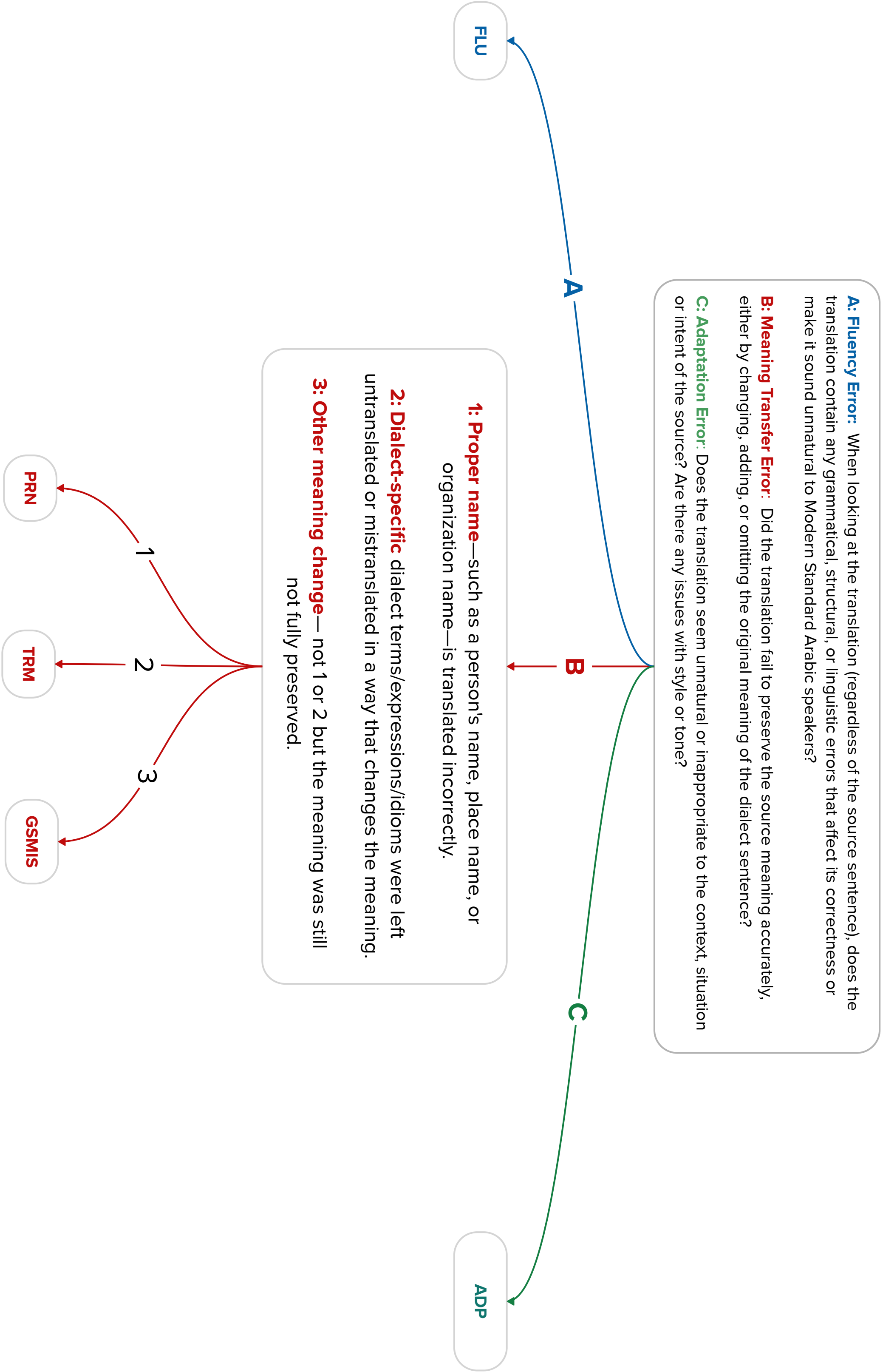
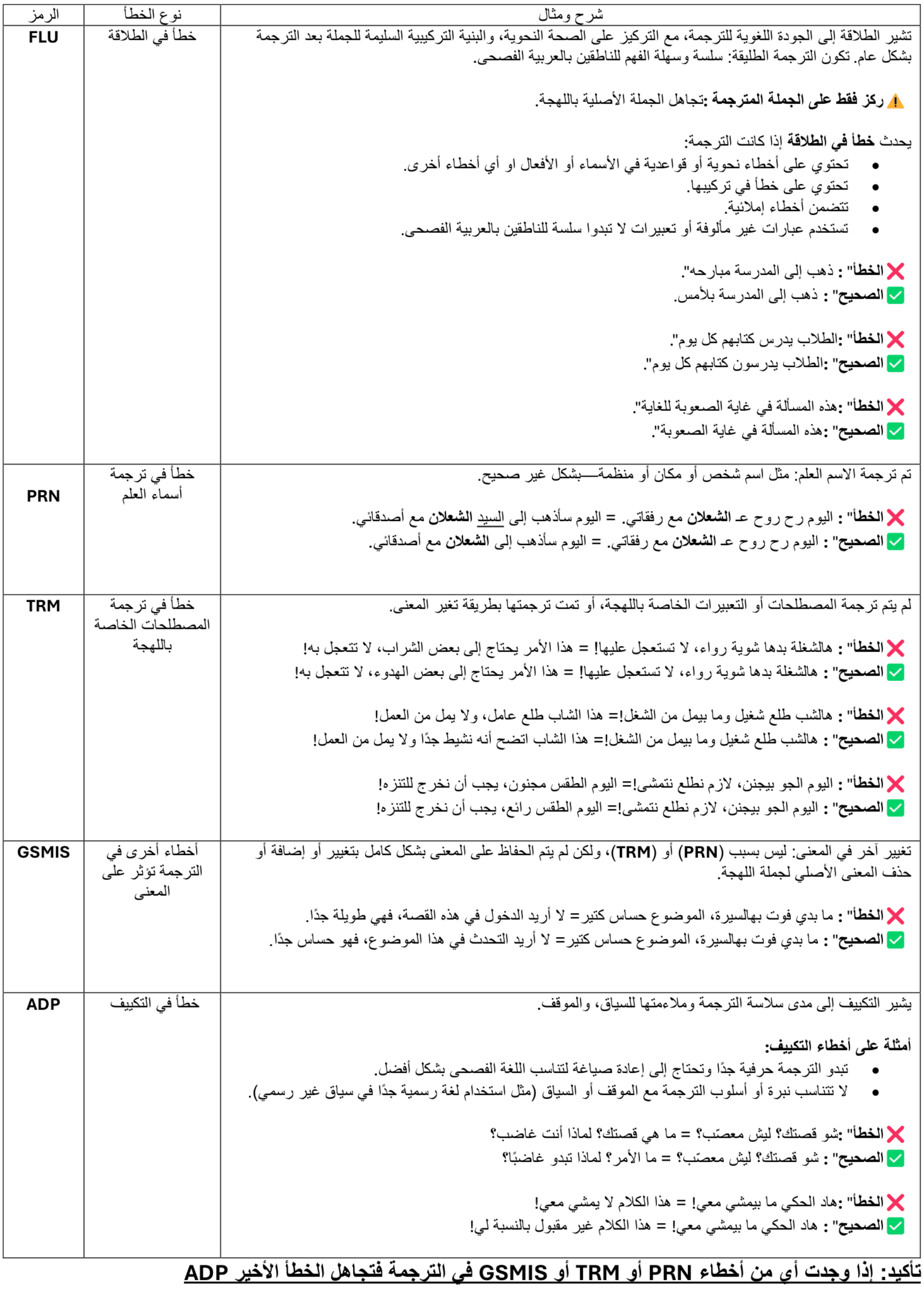

A: Fluency Error: When looking at the translation (regardless of the source sentence), does the translation contain any grammatical, structural, or linguistic errors that affect its correctness or make it sound unnatural to Modern Standard Arabic speakers?
B: Meaning Transfer Error: Did the translation fail to preserve the source meaning accurately, either by changing, adding, or omitting the original meaning of the dialect sentence?
C: Adaptation Error: Does the translation seem unnatural or inappropriate to the context, situation or intent of the source? Are there any issues with style or tone?
A
FLU
B
1: Proper name—such as a person's name, place name, or organization name—is translated incorrectly.
2: Dialect-specific dialect terms/expressions/idioms were left untranslated or mistranslated in a way that changes the meaning.
3: Other meaning change— not 1 or 2 but the meaning was still not fully preserved.
1
PRN
2
TRM
3
GSMIS
C
ADP

Figure 6: Structured decision tree guiding annotators through error classification for evaluating DA-MSA translations using the Ara-HOPE framework.



الرمز	نوع الخطأ	شرح ومثال
FLU	خطأ في الطلاقة	تشير الطلاقة إلى الجودة اللغوية للترجمة، مع التركيز على الصحة النحوية، والبنية التركيبية السليمة للجملة بعد الترجمة بشكل عام. تكون الترجمة الطليقة: سلسة وسهلة الفهم للناطقين بالعربية الفصحى. ⚠️ **ركز فقط على الجملة المترجمة :**تجاهل الجملة الأصلية باللهجة. يحدث **خطأ في الطلاقة** إذا كانت الترجمة: • تحتوي على أخطاء نحوية أو قواعدية في الأسماء أو الأفعال او أي أخطاء أخرى. • تحتوي على خطأ في تركيبها. • تتضمن أخطاء إملائية. • تستخدم عبارات غير مألوفة أو تعبيرات لا تبدوا سلسة للناطقين بالعربية الفصحى. ❌ **الخطأ" :** ذهب إلى المدرسة مبارحه". ✅ **الصحيح" :** ذهب إلى المدرسة بلأمس. ❌ **الخطأ" :**الطلاب يدرس كتابهم كل يوم". ✅ **الصحيح" :**الطلاب يدرسون كتابهم كل يوم". ❌ **الخطأ" :**هذه المسألة في غاية الصعوبة للغاية". ✅ **الصحيح" :**هذه المسألة في غاية الصعوبة".
PRN	خطأ في ترجمة أسماء العلم	تم ترجمة الاسم العلم: مثل اسم شخص أو مكان أو منظمة ـــــبشكل غير صحيح. ❌ **الخطأ" :** اليوم رح روح عـ **الشعلان** مع رفقاتي. = اليوم سأذهب إلى السيد **الشعلان** مع أصدقائي. ✅ **الصحيح" :** اليوم رح روح عـ **الشعلان** مع رفقاتي. = اليوم سأذهب إلى **الشعلان** مع أصدقائي.
TRM	خطأ في ترجمة المصطلحات الخاصة باللهجة	لم يتم ترجمة المصطلحات أو التعبيرات الخاصة باللهجة، أو تمت ترجمتها بطريقة تغير المعنى. ❌ **الخطأ" :** هالشغلة بدها شوية رواء، لا تستعجل عليها! = هذا الأمر يحتاج إلى بعض الشراب، لا تتعجل به! ✅ **الصحيح" :** هالشغلة بدها شوية رواء، لا تستعجل عليها! = هذا الأمر يحتاج إلى بعض الهدوء، لا تتعجل به! ❌ **الخطأ" :** هالشب طلع شغيل وما بيمل من الشغل!= هذا الشاب طلع عامل، ولا يمل من العمل! ✅ **الصحيح" :** هالشب طلع شغيل وما بيمل من الشغل!= هذا الشاب اتضح أنه نشيط جدًا ولا يمل من العمل! ❌ **الخطأ" :** اليوم الجو بيجنن، لازم نطلع نتمشى!= اليوم الطقس مجنون، يجب أن نخرج للتنزه! ✅ **الصحيح" :** اليوم الجو بيجنن، لازم نطلع نتمشى!= اليوم الطقس رائع، يجب أن نخرج للتنزه!
GSMIS	أخطاء أخرى في الترجمة تؤثر على المعنى	تغيير آخر في المعنى: ليس بسبب **(PRN)** أو **(TRM)**، ولكن لم يتم الحفاظ على المعنى بشكل كامل بتغيير أو إضافة أو حذف المعنى الأصلي لجملة اللهجة. ❌ **الخطأ" :** ما بدي فوت بهالسيرة، الموضوع حساس كتير= لا أريد الدخول في هذه القصة، فهي طويلة جدًا. ✅ **الصحيح" :** ما بدي فوت بهالسيرة، الموضوع حساس كتير= لا أريد التحدث في هذا الموضوع، فهو حساس جدًا.
ADP	خطأ في التكييف	يشير التكييف إلى مدى سلاسة الترجمة وملاءمتها للسياق، والموقف. **أمثلة على أخطاء التكييف:** • تبدو الترجمة حرفية جدًا وتحتاج إلى إعادة صياغة لتناسب اللغة الفصحى بشكل أفضل. • لا تتناسب نبرة أو أسلوب الترجمة مع الموقف أو السياق (مثل استخدام لغة رسمية جدًا في سياق غير رسمي). ❌ **الخطأ" :**شو قصتك؟ ليش معصّب؟ = ما هي قصتك؟ لماذا أنت غاضب؟ ✅ **الصحيح" :** شو قصتك؟ ليش معصّب؟ = ما الأمر؟ لماذا تبدو غاضبًا؟ ❌ **الخطأ" :**هاد الحكي ما بيمشي معي! = هذا الكلام لا يمشي معي! ✅ **الصحيح" :** هاد الحكي ما بيمشي معي! = هذا الكلام غير مقبول بالنسبة لي!

تأكيد: إذا وجدت أي من أخطاء PRN أو TRM أو GSMIS في الترجمة فتجاهل الخطأ الأخير ADP

Figure 7: The annotation guidelines provided to human annotators explain each error type with illustrative examples, assisting them in using the Ara-HOPE Annotation Decision Tree.