Multilingual Dialogue Generation and Localization with Dialogue Act Scripting

Anonymous ACL submission

Abstract

Non-English dialogue datasets are scarce, and models are often trained or evaluated on translations of English-language dialogues, an approach which can introduce artifacts that reduce their naturalness and cultural appropriateness. This work proposes Dialogue Act Script (DAS), a structured framework for encoding, localizing, and generating multilingual dialogues from abstract intent representations. Rather than translating dialogue utterances directly, DAS enables the generation of new dialogues in the target language that are culturally and contextually appropriate. By using structured dialogue act representations, DAS supports flexible localization across languages, mitigating translationese and enabling more fluent, naturalistic conversations. Human evaluations across Italian, German, and Chinese show that DAS-generated dialogues consistently outperform those produced by both machine and human translators on measures of cultural relevance, coherence, and situational appropriateness.[1]

1 Introduction

Developing multilingual dialogue systems requires high-quality conversational data across diverse languages. However, authentic dialogue datasets are often scarce, costly, or difficult to obtain, making it challenging to train robust multilingual models (Casanueva et al., 2022). A common technique is to generate synthetic dialogues by translating existing English dataset, but this approach often fails to capture cultural nuances and conversational norms leading to two key issues: anglocentric biases, the assumption that English-speaking cultural contexts are universally applicable, and artifacts that make dialogues sound unnatural in the target language (Artetxe et al., 2020).

For instance, dialogues translated from English may retain American or British settings, mention culturally specific brands, or use names common in English-speaking countries but rare elsewhere. These issues may leave the dataset culturally English limiting its usefulness for training and evaluating linguistically and culturally diverse dialogue systems.

To overcome these limitations, previous work has explored outline-based dialogue generation, where structured prompts rather than full English dialogues guide the creation of new conversational data (Shah et al., 2018; Majewska et al., 2023). Majewska et al. (2023) showed that this approach produces more natural and culturally appropriate dialogues than translations by professional human translators, as native speakers prefer localized adaptation over direct translation. However, their method relied on human annotators, limiting its scalability.

Building on this idea, we propose Dialogue Act Script (DAS), a structured framework for encoding, localizing, and generating multilingual dialogues. By abstracting conversations into intent-based representations before localization, DAS enables scalable, automatic adaptation of dialogue content while avoiding both anglocentric biases and translationese. This approach retains the strengths of outline-based annotation while leveraging large language models (LLMs) for both abstraction and localization, producing natural and culturally appropriate dialogues without requiring human annotation.

This work investigates the following research questions:

1. How does representing dialogues with DAS affect the fluency, coherence, and cultural appropriateness of generated dialogues across languages?

2. To what extent does DAS enable greater interpretability and control in multilingual dialogue generation?

[1] Code and data to be released upon acceptance.

3. What are the trade-offs between structured and flexible intent representations in DAS, and how do they affect reproducibility and dialogue quality?

4. How reliably can automated evaluation methods using large language models approximate human judgments of dialogue quality?

By addressing these questions, we aim to demonstrate that Dialogue Act Scripting (DAS) facilitates more culturally appropriate and coherent multilingual dialogue generation, as evaluated through both automated and human assessments across multiple languages.

To evaluate our approach, we use XDailyDialog (Liu et al., 2023) and the Cross-lingual Outline-based Dialogue (COD) dataset (Majewska et al., 2023). We compare DAS-generated dialogues against both machine-translated and human-translated versions of the original English dialogues. While translation is the most common method for building multilingual dialogue corpora, our results show that DAS-generated dialogues consistently outperform translation-based baselines on human evaluations.

2 Related Work

Translation-based methods are a common strategy for creating multilingual dialogue datasets (Mendonca et al., 2023; Anastasiou et al., 2022; Lin et al., 2021; Liu et al., 2023), but they often introduce structural inconsistencies that affect model generalization. Artetxe et al. (2020) show that translated datasets fail to reflect naturally occurring multilingual data due to translation artifacts that distort linguistic patterns. These distortions can lead to unnatural exchanges and discourse inconsistencies, limiting their effectiveness for training conversational models.

To mitigate these issues, human-guided annotation methods have been explored. Majewska et al. (2023) introduced outline-based annotation, where annotators structure dialogues using prompts rather than full English translations. This approach enables cultural adaptation and prevents artificial alignment, leading to more natural multilingual dialogues. While effective, manual annotation is resource-intensive and difficult to scale.

An alternative is synthetic dialogue generation, where models generate dialogues autonomously. Shah et al. (2018) introduced Machines Talking to Machines (M2M) to generate large-scale synthetic dialogues, but such methods risk producing artificial conversational patterns that diverge from human interactions.

Recent work has explored how LLMs can generate structured representations from natural language. Li et al. (2023) turned information extraction into a code generation task, using Code-LLMs to produce structured outputs. Similarly, Sainz et al. (2024) introduced GoLLIE, a guideline-aware LLM for zero-shot IE, which uses annotation guidelines structured as Python classes to improve IE accuracy. These approaches show that LLMs can effectively generate structured, code-like representations as well as free-form text.

3 Dialogue Act Script

3.1 Overview

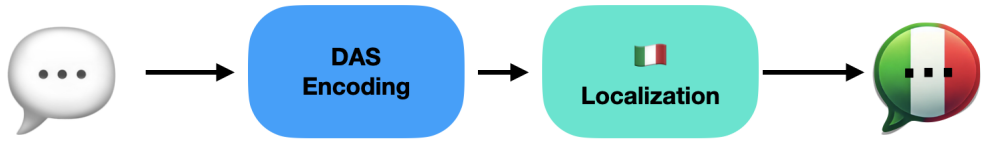
DAS is a structured framework for encoding dialogue through functional abstraction. It represents communicative intent using a predefined set of dialogue acts and parameters. Dialogue acts categorize utterances based on their communicative function (e.g., requesting, informing, or directing) rather than their surface form (Austin, 1962).

Rather than preserving the surface form of source-language dialogues through direct translation, DAS enables culturally adaptive generation by abstracting dialogues into structured intent representations and regenerating them in the target language. This approach helps mitigate anglocentric biases, reduces artifacts associated with literal translation, and supports the creation of more natural, contextually appropriate dialogues across languages.

3.2 DAS Pipeline for Multilingual Dialogue Generation

DAS facilitates the creation of multilingual dialogue data by culturally adapting dialogues through a three-step process, as illustrated in Figure 1:

Encoding: Each utterance is converted into a DAS representation by classifying its dialogue act and extracting only the essential components needed to preserve its intent and function, such as the speaker, action, relevant conditions, and timeframe. This structured abstraction preserves communicative intent while allowing for flexible multilingual reconstruction. For example, the English utterance “Actually, we are famous for our Cuervo Gold margaritas” may be



Do you have any house specials?	inquire(topic=menu, subject=house_specials, availability=yes_no)	inquire(topic=menu, subject=house_specials, availability=yes_no)	Avete delle specialità della casa?
Actually, we are famous for our Cuervo Gold margaritas.	inform(subject=restaurant, attribute=famous, object=**Cuervo_Gold_margaritas**)	inform(subject=restaurant, attribute=famous, object=**Negroni**)	Siamo famosi per il nostro Negroni.
That sounds good! Please bring me one of those.	express(approval); seek_action(action=bring, object=**Cuervo_Gold_margarita**)	express(approval); seek_action(action=bring, object=**Negroni**)	Ah, il Negroni! Sembra una buona scelta. Potrebbe portarmene uno, per favore?
Would you like that drink blended or on the rocks?	inquire(topic=drink_preference, subject=**Cuervo_Gold_margarita**, options=**[blended, on_the_rocks]**)	inquire(topic=drink_preference, subject=**Negroni**, options=**[with_ice, without_ice]**)	Certo! Lo preferisce con più ghiaccio o senza?

Figure 1: The DAS localization pipeline

encoded as `inform(subject=restaurant, attribute=famous, object=Cuervo_Gold_margaritas)`.

Localization: The DAS representation is then adapted to align with cultural norms in the target language by modifying relevant parameters (e.g., named entities, cultural references, or commonly used objects) while preserving the original dialogue act and intent. For instance, when adapting for an Italian audience, the drink `Cuervo_Gold_margaritas` might change to `Negroni`, reflecting a more common cocktail in Italian bars.

Decoding: Finally, the localized DAS representation is realized as fluent, coherent dialogue in the target language. This generation step reconstructs the conversation in a culturally appropriate manner while remaining faithful to the original communicative intent. For example, the localized representation `inform(subject=restaurant, attribute=famous, object=Negroni)` could be decoded into Italian as: Siamo famosi per il nostro Negroni. ("We are famous for our Negroni")

3.3 Encoding

The encoding process separates the form and content of an utterance, producing a structured representation that captures intent, dialogue acts, and semantic roles. This step consists of three key components:

Dialogue Act Classification: Each utterance is assigned a dialogue act representing its communicative function (e.g., requesting, informing, expressing). This abstraction captures speaker intent independently of linguistic form, ensuring consistent representation across languages and phrasing styles. DAS is agnostic to the specific taxonomy used; any consistent set of communicative functions can be employed, or even omitted entirely in more free-form representations (See Appendix E for experiments). In this study, we use a custom taxonomy developed to balance coverage, annotator consistency, and generation utility (see Section 4.1).

Slot Filling/Semantic Role Labeling: Key roles and entities are assigned to fill the parameters of the dialogue acts. These parameters provide the minimum necessary information to reconstruct the utterance while preserving intent. This structured format ensures that critical details are explicitly captured, facilitating accurate localization and natural dialogue generation. For example, the utterance "The wine list is on the second page of your menu." can be represented as: `inform(subject=wine_list, location=second_page, object=menu)` This representation captures the essential meaning while abstracting away language-specific phrasing, allowing for more flexible adaptation across different languages and cultural contexts.

Speaker Identification: To maintain conversational coherence, each utterance is labeled with speaker roles. Speakers are typically identified as "Speaker 1" and "Speaker 2," but when specific roles (e.g., "Student" and "Teacher") or named entities ("Susan" or "Billy") are present, they are retained to enhance dialogue flow.

To capture broader conversational context, we prompted the model to generate scenarios with character biographies, allowing for greater consistency in tone and formality. These biographies included details such as names, ages, genders, and relationships between speakers to ground the dialogue in a more natural setting. Further details, including the full prompt and ablation studies, are provided in Appendix D.

3.4 Localization

The localization step in DAS promotes cultural adaptability by enabling the generation of dialogues that are appropriate for the norms, entities, and expectations of the target language and culture. Rather than relying on direct translation from English, DAS supports the creation of multilingual datasets that avoid anglocentric biases by localizing from an abstract representation. At the same time, DAS offers flexibility: developers can choose more literal or more culturally adapted realizations depending on the application, enabling either near-equivalent phrasing or broader contextual adjustment.

In our implementation, localization is performed automatically by prompting a large language model to first adapt the contextual frame (e.g., names, locations, and cultural references), and then update individual DAS turns by modifying relevant parameters (e.g., `location=New York` → `location=Beijing`) while preserving the underlying dialogue act. This allows the communicative function to remain consistent while the realization reflects culturally relevant details.

3.5 Decoding

Decoding involves generating natural-language dialogue from the DAS representation. Given the character descriptions and setting, which may have been localized, each DAS turn is realized as a fluent, contextually appropriate utterance in the target language. This step also allows for fine-grained control over linguistic features. For example, developers can adjust the complexity or formality of the output to suit different audiences or use cases. A single DAS encoding such as `inquire(topic=menu, subject=house_specials)` might be decoded with simple grammar and vocabulary ("Do you have house specials?"), or as a more formal version ("Would you be able to tell me about the house specials currently on offer?") This flexibility makes DAS particularly useful for applications such as language learning.

Decoding can be performed turn-by-turn (e.g., in interactive chatbot settings) or over the entire dialogue (e.g., for full-script localization). The approach is language-agnostic: once localized, a DAS representation can be realized in any language supported by the generation model. In our experiments, we evaluate decoding across Chinese, Italian, German, and English to assess DAS's support for both cross-lingual and controlled-generation scenarios.

4 Experiments

To evaluate the effectiveness and flexibility of DAS, we conduct four experiments aligned with our research questions:

RQ1: Can LLMs reliably encode conversations into DAS representations? **RQ2:** Does the DAS representation preserve core meaning while allowing form variation? **RQ3:** How do slot-based localizations compare to human annotated localizations? **RQ4:** Can DAS localization produce dialogues that are more culturally relevant than direct translation? **RQ5:** Does the modular DAS pipeline offer advantages over end-to-end prompting?

For these experiments, we selected 50 dialogues from the DailyDialog dataset (Li et al., 2017), which covers a range of conversational topics, lengths, and emotional tones. To ensure a representative sample for translation and human evaluation, we applied the following criteria:

1. **Conversation Length**: Dialogues with 8 to 16 turns were selected, resulting in an average of 10.92 turns per dialogue.

2. **Topic Variety**: DailyDialog categorizes conversations into 10 distinct topics: Ordinary Life, School Life, Culture & Education, Attitude & Emotion, Relationship, Tourism, Health, Work, Politics, and Finance. We randomly selected 5 dialogues per topic to ensure diverse conversational contexts.

We use the XDailyDialog dataset (Liu et al., 2023) as a reference for professionally-translated

Annotator	Human1	Human2	GPT4o-mini
Human2	0.844	-	-
GPT4o-mini	0.765	0.746	-
GPT4o	0.822	0.769	0.805

Table 1: Inter-annotator agreement (Cohen's kappa) results for DAS function annotation.

dialogues in Italian, German, and Chinese. We also include a simple machine translation baseline by prompting GPT-4o to translate directly from English (see Appendix G.2 for the prompt).

While DAS is flexible and can be applied with different models at each stage, in this study, we use GPT-4o (gpt-4o-2024-08-06) and GPT-4o-mini (gpt-4o-mini-2024-07-18) for encoding, localization, and decoding [2]. Temperature was set to 0 for encoding to ensure consistent DAS representations across runs, as variation in function labeling could affect reproducibility. For localization and decoding, a temperature of 0.2 was chosen to allow for natural variation in expression while still preserving core meaning.

4.1 RQ1 - Encoding Accuracy

To assess the reliability of DAS function annotations, we conducted an inter-annotator agreement (IAA) study comparing human-human consistency and human-GPT agreement for DAS function labeling. Two human annotators labeled 105 dialogue turns from five randomly selected conversations, using a predefined set of DAS functions. Rather than adopting an existing taxonomy, we designed a new, task-specific schema to test how well large language models could apply unfamiliar classification schemes. This choice also reduced the risk of data leakage, since widely used taxonomies may have been encountered during model training. Annotators received the same function definitions and examples as the language models, ensuring consistent guidelines[3]. We evaluated GPT-4o and GPT-4o-mini using identical propmts and instractions. The results are shown in Table 1.

High agreement between human annotators (κ = 0.844) suggests that the schema supports annotation consistency. Substantial agreement between humans and GPT-4o (κ = 0.822, 0.769) indicates that the LLM can reliably apply dialogue acts when provided with clear definitions and examples. GPT-4o-mini also maintained reasonable agreement ($\kappa = 0.765, 0.746$), though slightly lower than GPT-4o.

To assess the compatibility of DAS encoding with existing dialogue act schemes, we conducted an additional experiment using the DailyDialog taxonomy (Inform, Question, Directive, Commissive). GPT-4o was prompted to assign one of these four acts to each turn. GPT-4o achieved high F1 scores for Inform (0.92) and Question (0.94), which together covered 87.9% of all turns. Performance on the comparatively rarer Directive (0.63) and Commissive (0.64) was lower. This suggests that GPT-4o is strong at classifying more common and straightforward dialogue acts.

4.2 RQ2 - Encoding Meaning Preservation

To assess how well DAS preserves meaning while allowing for structural changes, we decoded DAS-encoded English dialogues back into English and compared them to the original dialogues. This evaluation serves two key purposes: first, to determine whether DAS retains the essential communicative intent of a conversation, and second, to examine whether DAS reconstruction introduces meaningful paraphrasing effects that could be useful for fluency enhancement or synthetic data generation.

We conducted human assessments using a pair of native English speakers. Annotators were shown pairs of conversations, the original dialogue and its DAS-decoded version, and asked the following questions:

1. Fluency: Which conversation has the more fluent or natural sounding language?
2. Coherence: Which conversation makes the most logical sense? (No sudden changes of topic, each turn naturally follows the previous on)
3. Situational Appropriateness: Which conversation has the more appropriate tone or style for the situation?
4. Meaning Preservation: How similar are the conversations in meaning?

For the first three questions annotators were allowed to choose, A, B, Both, or Neither. Win rates were calculated by assigning a point to a system each time it was chosen over another or when "Both" was selected; no points were awarded when

[2]GPT models were accessed through OpenAI's API and followed OpenAI's terms for API usage. The number of parameters of these models is undisclosed. We spend approximately $100 USD on experiments.

[3]The full taxonomy and examples are provided in Appendix A.

Metric	DAS	Original
Fluency	0.727	0.455
Logical Flow	1.000	0.636
Situational	0.909	0.636
Meaning Preservation	Avg. Score: 4.63/5	

Table 2: Human evaluation of DAS-decoded English compared to the original dialogues.

Language	Precision	Recall	F1
Arabic	0.929	0.760	0.836
Indonesian	0.865	0.802	0.832
Russian	0.894	0.796	0.843
Swahili	0.852	0.703	0.770
Average	0.885	0.765	0.820

Table 3: Slot-level comparison between GPT-localized and human-localized dialogues.

Language	Entertainment	Travel
Arabic	0.233	0.783
Indonesian	0.127	0.695
Russian	0.000	0.768
Swahili	0.008	0.713
Average	0.092	0.740

Table 4: Proportion of DAS-generated named entity localizations matching any human-annotated value, grouped by language and semantic category.

"Neither" was selected. Meaning preservation was reported on a Likert scale, with 1 indicating the conversations had completely different meanings, and 5 being they are identical in meaning.

The results, reported in Table 2, suggest that DAS decoding does not introduce many disfluencies or disrupt conversational flow. In most cases, DAS produces output that is at least as coherent and appropriate as the original dialogue, with notable improvements in fluency for over half of the conversations. The high meaning preservation score (4.63/5) indicates that DAS retains core intent effectively, even when rewording utterances. Although DAS generally improved fluency, situational appropriateness was slightly lower in some cases, suggesting that certain stylistic nuances may change during decoding.

In addition to human evaluation, we used automated metrics to assess the semantic similarity and structural differences between the original dialogues and their DAS-decoded versions. See Appendix B for details and results of this experiment.

4.3 RQ3 - Cultural Adaptation

To evaluate whether the DAS localization process produces culturally adapted slot substitutions similar to those made by human annotators, we conducted a slot-level comparison using dialogues from the Cross-lingual Outline-based Dialogue (COD) dataset (Majewska et al., 2023).

COD was created through manual rewriting of outlines, including a localization step where culturally specific named entities were replaced by native speakers. We applied DAS localization to the 92 original English dialogues from the COD development data and evaluate the 1196 annotated slots that contain values. First we look at how well DAS identifies slots that should be changed. We calculate the F1 score for localized slots using COD as the gold standard and report the results in Table 3.

To better understand how well DAS handles named entity localization, we grouped relevant slot types into two broad categories: Entertainment-related (e.g., Song Title, Actor, Director, Music Artist) and Travel-related (e.g., City Name, Airline). For each slot type in each language, we computed the proportion of DAS-generated localized slot values that matched any of the corresponding localized values selected by human annotators. We assume that values within a slot type are interchangeable if they fulfill similar cultural or geographic functions (e.g., New York → Jakarta or Bali). We then aggregated the instance-level matches to produce a category-level match rate (Table 4).

Travel-related slots, such as City Name and Airline, showed high match rates (87%-94%) across languages. These categories draw from a small, culturally salient set of entities, so making these categories show a more confident alignment with human localizations. In contrast, Entertainment-related slots (e.g., actors, directors, music artists) had lower match rates due to their open-ended nature. However, many of these apparent mismatches stem from the limitations of automatic evaluation, which only counts matches against a small set of annotated alternatives and may miss other valid substitutions.

To assess this, we manually annotated a sample of Indonesian outputs. The results revealed that the automatic method substantially underestimates true accuracy: for instance, Music Artist achieved 66.7% accuracy upon manual review, compared to just 4% under automatic matching. Overall, manual correction raised the match rate for the Entertainment category from 12.7% to 73.7%, suggesting that the performance of DAS in these categories is stronger than the automated metrics suggest.

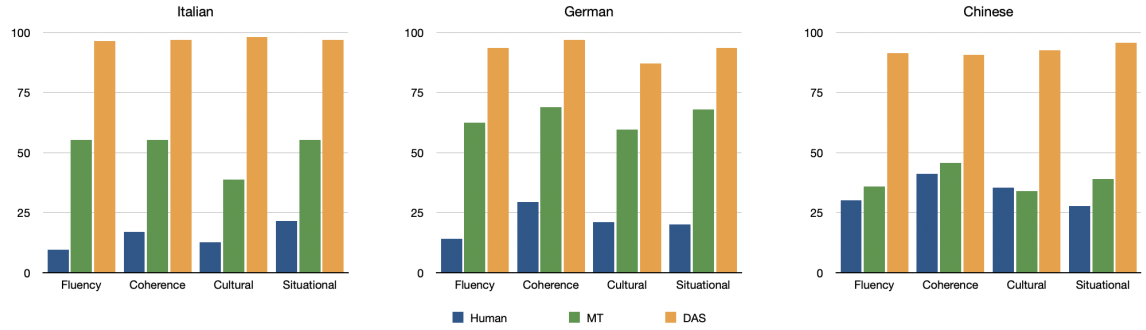


Figure 2: Win rates of each system across evaluation criteria (fluency, coherence, cultural relevance, and situational appropriateness). Higher win rates indicate stronger performance in pairwise comparisons.

4.4 RQ4 - Decoded Dialogue Quality

To assess the quality of dialogues generated through DAS localization, we conducted a human evaluation on the generated target-language text. We compare the DAS generated text to two different translation baselines, as translation is the common technique for generating multilingual datasets.

Two native speakers each of Chinese, Italian, and German were recuited to compare DAS-localized dialogues against two baselines: human-translated dialogues from the XDailyDialog dataset, and machine-translated dialogues generated by prompting GPT-4o to directly translate the English source. Although both baselines involve translation, we do not evaluate "translation accuracy"; instead, we treat these as standard approaches to multilingual dialogue generation and compare them to DAS as alternative generation methods. As we are not judging typical translation metrics such as fidelity to the source, we do not show the annotators the original English dialogues.

Annotators were presented with a random pair of generated dialogues and asked the following questions[4]:

1. Fluency: Which conversation has the more fluent or natural sounding language?

2. Coherence: Which conversation makes the most logical sense? (No sudden changes of topic, each turn naturally follows the previous on)

3. Cultural Relevance: Which conversation feels more culturally (Italian/German/Chinese)?

4. Situational Appropriateness: Which conversation has the more appropriate tone or style for the situation?

Each annotator was allowed to select A, B, Both, or Neither for each question. Win rates are calculated as in section 4.2.

The results, shown in Figure 2, demonstrate that DAS consistently outperforms or matches both machine translation and human translations, particularly in cultural relevance and situational appropriateness. To assess statistical significance, we conducted a two-tailed binomial test, comparing wins and losses only (excluding "Both" and "Neither" responses). Across all three languages and all four evaluation criteria, DAS was preferred over the professional translations with high significance ($p < 0.00001$).

While the lower performance of the professionally translated dialogues may seem surprising at first, these results may simply reflect a fundamental difference in goals between traditional translation workflows and open-ended, culturally adaptive dialogue generation.

Professional translators often aim to preserve the original meaning as faithfully as possible. However, as we saw in 4.2, the original dialogues contain disfluencies, inconsistent tenses, or informal phrasing, all of which could have lead to translations that feel rigid or unnatural in the target language. For example, one annotator noted that a professional translation shifted awkwardly between past and present tense, likely due to literal adherence to the original English. Such artifacts, while arguably accurate, are often dispreferred by native speakers evaluating fluency and conversational naturalness.

In contrast, GPT-4o, even under a simple translation prompt, tends to "clean up" awkward or incon-

[4]Questions were translated into the target language using GPT-4o.

Metric	DAS	Single Prompt
Fluency	0.78	0.26
Logical Flow	0.83	0.45
Cultural	0.80	0.32
Situational	0.78	0.25

Table 5: Win-rates of DAS compared to the single prompt translate-localize averaged across all languages.

sistent source material during generation, resulting in smoother target-language output. DAS goes a step further by discarding the surface form of the source entirely. Its reliance on abstract, intent-based representations allows for even greater flexibility in how conversations are realized, enabling shifts in style, tone, and cultural framing that better align with local conversational norms.

It is also important to consider the nature of the evaluation setup as a pairwise comparison instead of quality scores. As such, the fact that professional translations were often dispreferred does not imply that they are low-quality. Instead, it reflects their performance relative to more adaptive systems in a specific conversational context.

These findings align with those reported by Majewska et al. (2023), who similarly observed that dialogue outputs generated from abstract representations were preferred over direct translations. Together, these results suggest that abstraction-based pipelines like DAS may be more effective than form-preserving translation approaches when the goal is to generate fluent, culturally appropriate dialogue, rather than to maintain strict fidelity to source-language wording.

4.5 RQ5 - DAS Pipeline vs. Single Prompt

Since the DAS pipeline currently relies on GPT-4o for all three steps, a natural question arises: could a single prompt accomplish the same task more efficiently? To test this, we constructed a baseline that prompts GPT-4o to directly translate and localize the English dialogue into the target language in one step. This prompt uses the localization instruction used in the DAS pipeline but skips the intermediate abstraction step entirely.

As shown in Table 5, despite receiving the same high-level localization instructions, the single-prompt baseline was consistently dispreferred across all evaluation criteria. Human annotators noted several recurring issues with the single-prompt approach. In many cases, cultural localization was incomplete or entirely absent. For example, in the case shown in Figure 1, references to "Cuervo Gold margaritas" were preserved verbatim rather than adapted to locally appropriate alternatives. Annotators also reported that the single-prompt outputs tended to sound "textbook-like" or sometimes inappropriately casual or formal. In particular, one Italian annotator described the style as stiff and lacking conversational naturalness.

These results demonstrate that the performance gains observed with DAS are not solely due to the use of GPT-4o, but emerge from the modular pipeline itself. Explicitly separating the localization and decoding steps appears to improve both cultural relevance and fluency, even when using the same base model.

5 Conclusion

This study introduced Dialogue Act Script, a modular framework for abstracting and localizing multilingual dialogues through intent-based representations.By separating the processes of encoding, localization, and decoding, DAS enables explicit cultural adaptation and flexible realization of dialogue across languages.

In our experiments, DAS-based translations consistently outperformed both human and machine translations. As shown in Section 4.5, these gains reflect the benefits of modular design: separating communicative intent from surface form enables more flexible and culturally adaptive generation, independent of any single model like GPT-4o.

A central strength of DAS is its modularity. Each step in the pipeline is independent, allowing for greater adaptability. While this paper used GPT-4o for all stages, there is growing evidence of cultural and stylistic biases in LLMs, including anglocentric tendencies and uneven performance across languages (Naous et al., 2024). DAS makes it possible to substitute any component with an alternative model, a retrieval-based method, or a human-in-the-loop process. Exploring these modular configurations is a promising direction for future work.

Beyond localization, DAS presents new opportunities for synthetic data generation, multilingual AI training, and rule-based machine translation in low-resource settings. We leave addressing challenges such as annotation consistency, scalability, and domain adaptability to future work.

Limitations

Several limitations apply to the current version of this work. First, the DAS pipeline relies on multiple

calls to LLMs, which increases computational cost. Although the DAS encoding step is reusable across languages, deploying the pipeline in low-resource or compute-constrained environments remains challenging. Future work should explore lighter-weight or retrieval-based alternatives for each step of the pipeline, especially for localization and decoding.

Second, our human evaluation is limited to the XDailyDialog dataset, which consists of open-domain chitchat dialogues. While this setting is useful for evaluating conversational fluency and cultural adaptation, it does not represent the structure or communicative goals of more specialized domains. Future work should explore how well DAS generalizes to task-oriented or domain-specific dialogues, such as those found in customer support, healthcare, or legal contexts.

Third, while DAS is designed to enable cultural adaptation, the current implementation relies entirely on GPT-4o for all steps of the pipeline. This raises valid concerns about inherited cultural biases from the underlying model, particularly given prior findings on anglocentric bias in LLMs (Naous et al., 2024). Our intention is not to claim that GPT-4o is an ideal solution for localization, but rather to evaluate whether DAS, as an abstraction framework, enables more flexible and culturally responsive generation than translation alone. DAS is modular by design: each step can be implemented independently. The localization step, in particular, does not require generation and could be replaced with rule-based substitutions, retrieval systems, or human annotations. We see improving the localization step as an important direction for future work.

Fourth, while we evaluate the end-to-end quality of localized dialogues through pairwise human judgments, we do not directly validate the cultural appropriateness of individual slot substitutions. A more targeted evaluation of the localization step, for instance through native speaker judgments of entity familiarity or cultural fit, remains an important area for future study.

Finally, our evaluation primarily targets well-resourced languages such as Chinese, Italian, and German. The performance of DAS in low-resource or morphologically complex languages remains uncertain. Although we include slot-level analysis for additional languages in the COD dataset, further work is needed to understand how DAS performs in settings where LLMs have limited coverage or cultural knowledge.

Ethical Considerations

We recruited human annotations for evaluating DAS-generated dialogues, including two native speakers each for German, Chinese, and English, one contributing author for English, and one for Indonesian. All annotators participated voluntarily and offered compensation of $10–$15 USD per hour depending on location. Annotators were informed of the task scope and consented to participate under conditions aligned with ethical research practices.

As with all work involving LLMs, our framework inherits risks related to unintended social and cultural biases. One recurrent pattern was a default tendency to assign male-female gender roles to dialogue participants, with 88% of conversations exhibiting this distribution. Although some mitigation strategies were attempted, this bias persisted. We did not conduct an exhaustive analysis of other cultural or representational biases, particularly in localized content. Future work should include more targeted bias evaluation and mitigation strategies, and we caution users of DAS to critically assess outputs, especially in real-world or sensitive applications.

The use of LLMs in our pipeline contributes to the environmental footprint of large-scale NLP systems. Future work could explore lightweight models or optimization strategies to improve the sustainability of multilingual generation frameworks like DAS.

We use the XDailyDialog dataset under the Apache-2.0 License, and its base dataset, DailyDialog, under CC BY-NC-SA 4.0. Both licenses permit research use with attribution. The original English conversations were sourced from websites for English learners and primarily reflect informal chitchat dialogues, which may not generalize to other conversational domains.

While DAS supports cultural adaptation of dialogues, it is not designed for high-stakes applications such as legal, medical, or financial translation. Any deployment beyond research settings should include human validation and safeguards to ensure responsible use.

AI tools such as ChatGPT and GitHub Copilot were used for minor language revisions and line-level code assistance, but all research design and outputs were authored and verified by the research team.

References

Dimitra Anastasiou, Anders Ruge, Radu Ion, Svetlana Segărceanu, George Suciu, Olivier Pedretti, Patrick Gratz, and Hoorieh Afkari. 2022. A machine translation-powered chatbot for public administration. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 329–330, Ghent, Belgium. European Association for Machine Translation.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

John Langshaw Austin. 1962. *How to do things with words*. William James Lectures. Oxford University Press.

Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. 2022. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013, Seattle, United States. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. CodeIE: Large code generation models are better few-shot information extractors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Toronto, Canada. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021. XPersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112, Online. Association for Computational Linguistics.

Zeming Liu, Ping Nie, Jie Cai, Haifeng Wang, Zheng-Yu Niu, Peng Zhang, Mrinmaya Sachan, and Kaiping Peng. 2023. XDailyDialog: A multilingual parallel dialogue corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12240–12253, Toronto, Canada. Association for Computational Linguistics.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2023. Towards multilingual automatic open-domain dialogue evaluation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 130–141, Prague, Czechia. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. *Preprint*, arXiv:2310.03668.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 3 (Industry Papers), pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A DAS Functions

1. **Inquire**

   Seeks information or clarification. Includes direct questions or indirect inquiries.

   What time does the meeting start?

2. **Clarify**

   Seeks to resolve ambiguity, misunderstanding, or confusion in a previous statement. Often involves rephrasing, elaboration, or highlighting specific details.

   I meant next Tuesday.

3. **Inform**

   Provides factual information, details, or observations.

   This policy was updated last week.

4. **Express**

   Communicates emotions, attitudes, or subjective opinions.

   That's an excellent idea!

5. **Agree**

   Affirms or aligns with a previous statement.

   Yeah, that makes sense to me.

6. **Disagree**

   Explicitly communicates disagreement or contradiction with a previous statement or idea. May provide reasoning or counterarguments but does not necessarily imply hostility or conflict.

   That doesn't seem right to me.

7. **Commit**

   Explicitly agrees or promises to take a future action, either in response to a request or as a declaration of intent. The action must be something the speaker is directly responsible for performing.

   Yes, I'll take care of that.

8. **Acknowledge**

   Neutral receipt of information, often used for backchanneling or minimal responses.

   I see.

   Okay.

9. **Seek Action**

   Represents any utterance where the speaker seeks to influence the listener's behavior, encompassing both polite requests and authoritative commands.

   Could you please send me the file?

   Turn off the light.

10. **Suggest**

    Proposes an action, idea, or alternative. May include advice or recommendations.

    Why don't you try restarting your computer?

11. **Offer**

    Voluntarily provides help, solutions, or resources.

    Would you like some water?

12. **Reject**

    Declines or refuses a proposal, offer, or request. May provide justification or explanation, though this is not required.

    I'm sorry, but I'll have to pass.

13. **Encourage**

    Provides motivation, praise, or positive reinforcement.

    Don't worry, you'll figure it out!

14. **Manage Topic**

    Handles transitions between conversation topics. Can be used for opening, changing, or closing topics.

    Let's move on to the next point.

15. **Social Interaction**

    Includes greetings and meaningless small talk designed for polite social interaction.

    Hello.

    How are you?

    Fine. And you?

B Automated Evaluation of Decoding Back into English

We evaluated DAS-decoded English using GPT-4o and GPT-4o-mini, and a direct paraphrase baseline, where the original dialogues were rephrased using a simple paraphrasing prompt[5]. The paraphrase baseline provides a useful reference point for distinguishing ordinary surface rewording from the more structured transformations introduced by DAS. For example, given the original utterance, "I'm a bit worried about you going shopping by yourself this afternoon." the paraphrased baseline produces "I'm a little concerned about you heading out to shop alone this afternoon." In contrast, DAS decoding generates "I'm a bit worried about you going shopping alone. Are you sure you'll be okay?" While the paraphrase baseline makes minor lexical and syntactic adjustments, DAS introduces a more structured transformation by breaking the utterance into multiple turns, adding conversational nuance, or adjusting for different dialogue dynamics.

To ensure robustness and consistency, each model was tested across three runs with a temperature setting of 0.2. To mitigate potential biases, we fixed the encoder and varied the LLM used for DAS decoding, allowing us to assess the effect of different decoding strategies in DAS. The reported scores represent the averages across all runs.

For automated evaluation, we computed BERTScore (Zhang et al., 2019) to measure meaning retention, BLEU (Papineni et al., 2002) to quantify lexical overlap, and ChrF++ (Popović, 2015) to evaluate character-level and word-level similarity between the original and DAS-decoded texts. Since DAS does not use the original sentence as input, we expect the BLEU score to be lower than paraphrasing, while the BERTScore remains high. ChrF++ captures both word- and character-level overlap, making it more flexible than BLEU in handling reworded outputs. However, since DAS modifies sentence structure more than standard paraphrasing, we still expect ChrF++ scores to be lower than paraphrasing reflecting content preservation despite structural variation. The results are summarized in Table 6.

The lower BLEU scores compared to the paraphrase baseline suggest that DAS decoding introduces lexical variety, making it distinct from simple word-for-word reformulation. The ChrF++ scores also show that DAS reformulations diverge more from the original structure than direct paraphrasing. Despite this increased divergence, BERTScore remains high (over 0.9, even for the smaller system), reinforcing that DAS effectively preserves intent while rewording the dialogue more flexibly than standard paraphrasing. The fact that DAS decoding does not have direct access to the original sentence yet still scores relatively close to the paraphrase baseline suggests that its structured encoding influences realization in ways that may limit extreme rewording. Future work could explore whether adjusting encoding constraints allows for more diverse yet meaning-preserving reformulations.

[5]See Appendix G.1

Model	BERTScore	BLEU	ChrF++
Paraphrasing	0.943	0.184	0.389
GPT4o-mini	0.909	0.126	0.343
GPT4o	0.914	0.142	0.369

Table 6: Semantic (BERTScore) and form-focused (BLEU/ChrF++) similarities between the original and the decoded utterances

C Automated Evaluation of Localization Quality

Human evaluation is not always available or practical at scale, particularly for multilingual dialogue assessment, where hiring expert annotators for every language is costly and time-consuming. To determine whether GPT-4o can serve as a reliable evaluation tool, we tested its ability to judge conversation quality using the same criteria as human annotators.

We prompted GPT-4o with the same questions used in the human evaluation, one at a time, covering fluency, coherence, cultural relevance, and situational appropriateness. Each pair of translations was shown twice, with the order reversed in the second presentation to control for positional bias. The final annotation was determined by merging the two judgments: If GPT-4o selected the same conversation in both orders, it was counted as a win for that system, while conflicting responses were recorded as a tie.

To evaluate how well GPT-4o's judgments align with human preferences, we computed Cohen's Kappa between GPT-4o and the human annotators, both overall and for each evaluation metric individually. The human annotator judgment was aggregated using majority voting. The results are reported in Table 7.

The results indicate strong alignment between

Aspect	Italian	German	Chinese
Fluency	0.396	0.846	0.698
Coherence	0.287	0.610	0.795
Cultural Relevance	0.348	0.844	1.000
Situational Appropriateness	0.341	0.582	0.894
Overall	0.346	0.726	0.843

Table 7: Cohen's Kappa between GPT-4o and human annotators. For Italian and German, human annotations were aggregated using the majority vote of all annotators. For Chinese, a single native annotator was used.

GPT-4o and human judgments in some areas, particularly in cultural relevance and fluency for German and Chinese. This suggests that GPT-4o applies consistent evaluation criteria and broadly captures human preferences in some settings.

However, agreement varies across languages, with weaker alignment in Italian compared to German and Chinese. Situational appropriateness and coherence exhibit lower agreement for Italian and German, while fluency is more challenging for Chinese. These findings suggest that GPT-4o may struggle with contextual nuances in evaluation, and its reliability as an evaluator depends on both the target language and the specific quality dimension being assessed.

These findings suggest that GPT-4o can serve as a structured, scalable evaluation tool when large-scale human annotation is infeasible. However, language-specific inconsistencies must be considered. While alignment is strong in some cases, discrepancies in others highlight the need for further investigation into how GPT-based evaluation models process different languages and cultural norms. Future work should explore why GPT-4o's evaluation accuracy varies across languages and whether prompting strategies or calibration techniques can improve cross-linguistic consistency.

D Conversational Context

Early experiments localized and decoded dialogues using DAS alone, without additional conversational context. However, manual inspection and consultation with native speakers revealed room for improvement, particularly in situational appropriateness. The generated dialogues often sounded too formal or stiff in contexts where a more natural or casual tone would have been expected.

One key observation was that nuances such as politeness levels were often lost in the encoding process. This was likely because DAS focuses on extracting content rather than form, whereas politeness and tone are often conveyed through structural and lexical choices rather than explicit meaning. To address this, we incorporated broader conversational context by prompting GPT-4o to generate a summary of the conversation, along with speaker names and biographical details.

Since many languages rely on grammatical gender, we asked GPT-4o to infer or assign speaker genders as part of the biographical information. However, in the initial test, every generated dialogue featured one male and one female character, indicating a bias toward binary gender pairings. To mitigate this, we explicitly modified the prompt to encourage greater diversity in gender assignments.

After this change, the resulting speaker distribution was: 88% male-female, 6% male-male (MM), 2% female-female, 4% non-binary-female. Interestingly, for one conversation, a non-binary character was changed into a male character during localization into German and Italian, while remaining non-binary in Chinese. No other characters had gender altered during localization.

Method	Fluency	Coherence	Culture	Situation
Italian				
Localized	73	70	76	74
+ Context	91	85	86	89
German				
Localized	82	76	72	76
+ Context	89	85	86	89
Chinese				
Localized	77	78	79	81
+ Context	82	80	90	93

Table 8: Win rates against machine translation and human translation for including a context summary or not.

The results in Table 8 reflect GPT-4o-based evaluation of localized dialogues with and without additional conversational context. While the inclusion of speaker biographies and conversational summaries led to higher GPT evaluation across all criteria, it is important to recognize that GPT-based evaluation may not always align with human judgment (See Appendix C).

To better understand this discrepancy, we conducted a small-scale human verification study for Italian, as it exhibited the lowest agreement between annotators and GPT evaluations in prior assessments. Native Italian speakers reviewed a sample of 10 conversations and confirmed GPT's evaluations, suggesting that the inclusion of context genuinely improved fluency, cultural relevance, and situational adaptation. However, given the limited sample size, further human evaluation is required

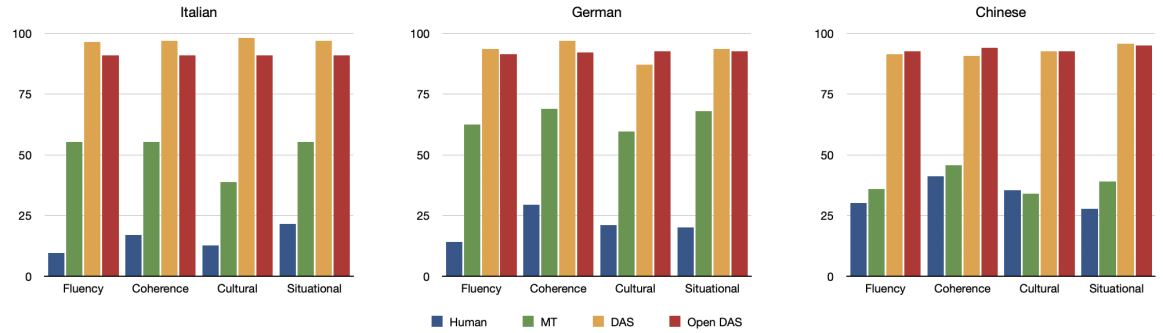


Figure 3: Win rates of each system, including Open DAS, across evaluation criteria.

to validate the extent of these improvements across different languages and conversational settings.

E Flexible Function Encoding with OpenDAS

The main variant of DAS used throughout this paper employs a structured function format, in which utterances are annotated with a predefined set of communicative functions. This constrained format supports consistency and reproducibility, and enables modular localization by allowing targeted changes to parameters while keeping the function label stable.

To explore a more expressive alternative, we introduce OpenDAS: a flexible encoding approach in which the model generates function labels freely, without being constrained to a fixed taxonomy. OpenDAS allows the LLM to define fine-grained communicative acts, potentially capturing more nuance in speaker intent.

The key difference between the structured DAS variant and OpenDAS lies in where the meaning is encoded. Structured DAS encodes most information in discrete parameters, while OpenDAS embeds more of the meaning directly into the function label. For example:

OpenDAS: `inquire_feelings_about_responsibility (responsibility=money)`

DAS: `inquire(topic=emotional_response, subject=responsibility, object=money, timeframe=current, aspect=feeling)`

This design difference has practical consequences. In the structured version, communicative intent is modular and easier to manipulate, such as swapping out specific cultural elements during localization. OpenDAS, by contrast, gives the model more explicit cues during decoding, which may aid fluency but constrain flexibility.

To quantify the impact of this difference on reproducibility, we computed inter-annotator agreement between human-annotated structured DAS functions and OpenDAS function labels generated by GPT-4o. As shown in Table 9, agreement dropped substantially under OpenDAS. This is unsurprising, as the model tends to create more specific or compound functions (e.g., `offer_help_with_booking`) that do not align with the coarser-grained categories used in human annotation. When these were truncated to include only the first word, agreement improved somewhat, suggesting the disagreement stems in part from the model introducing subtypes of functions.

Annotation Scheme	Human-GPT IAA
Closed DAS	0.822
Open DAS (Full)	0.080
Open DAS (Truncated)	0.269

Table 9: Inter-annotator agreement (Cohen's Kappa) for Closed DAS and Open DAS function annotation. Open DAS (Truncated) refers to cases where only the first word of the function label was considered.

Despite variability in labeling, OpenDAS performed comparably to the structured version in human evaluations[6]. As shown in Figure 3, preferences between OpenDAS and structured DAS varied slightly across languages, but no statistically significant differences were observed.

The results suggest that OpenDAS performs similarly to the structured version but does not consistently outperform it. While the taxonomy may not be strictly necessary for dialogue quality, it supports greater modularity and interpretability;

[6]Please note that this annotation was only conducted by a single annotator for each language.

particularly valuable if different systems are used for encoding, localization, and generation. We see OpenDAS as a viable alternative when simplicity is prioritized, though structured DAS offers stronger support for modular system design.

F Vector Embedding Analysis

To quantify the structural differences between the English and translated dialogues, we computed two embedding-based similarity metrics, each capturing a distinct aspect of linguistic variation:

- Cosine Similarity: Measures how closely the translated dialogue embeddings align with the English source. Lower values indicate greater syntactic and lexical divergence.
- KL-Divergence (Kullback and Leibler, 1951): Measures how much the probability distribution of translated embeddings diverges from that of the English source. Higher values indicate greater structural and lexical variability, reducing "translationese" effects.

All embeddings are computed using LaBSE (Language-Agnostic BERT Sentence Embeddings), a multilingual embedding model designed for cross-lingual similarity tasks (Feng et al., 2022). To assess whether translation methods differ significantly, we apply a one-way analysis of variance (ANOVA) for Cosine Similarity, which is expected to follow a normal distribution. For KL-Divergence, we use the non-parametric Kruskal-Wallis test (Kruskal and Wallis, 1952), which is more appropriate for non-normal distributions.

We evaluate three translation methods: Human Translation, which refers to the professional translations from XDailyDialog; Machine Translation, which consists of direct translations generated by GPT-4o; and DAS (ours), a translation approach implemented through DAS on top of GPT-4o. Table 10 presents the results of the analysis.

We analyze the structural and distributional shifts of DAS-generated dialogues compared to human and machine translations. ANOVA and Kruskal-Wallis tests confirmed statistically significant differences in cosine similarity ($F = 708.75$, $p < 0.0001$) and KL-Divergence ($H = 792.63$, $p < 0.0001$). These results indicate that DAS-generated dialogues exhibit significantly greater divergence from English sentence structures compared to both machine and human translations. Although human translations diverge more than machine translations, they still retain structural similarities. In contrast, DAS-generated dialogues exhibit even greater shifts, suggesting that they introduce more diverse sentence structures that better reflect target language norms.

Method	Cos Sim.	KL Div.
Italian		
Human	0.8254	0.0144
MT	0.9115	0.0064
DAS	0.6495	0.0303
German		
Human	0.8252	0.0144
MT	0.8992	0.0080
DAS	0.6549	0.0344
Chinese		
Human	0.8252	0.0144
MT	0.8741	0.0093
DAS	0.6794	0.0240

Table 10: Statistical analysis of cosine similarity (Cos Sim.) and KL-Divergence (KL Div.) between English source texts and their translations from XDailyDialog. ANOVA and Kruskal-Wallis tests confirm statistically significant differences ($p < 0.0001$).

KL-divergence results suggest that DAS produces more distributional variation, avoiding "translationese" effects common in machine-generated translations. This reinforces the potential of DAS to reduce anglocentric biases in multilingual dialogue generation by encouraging more natural and varied sentence structures.

These findings suggest that DAS may be particularly useful for multilingual dialogue systems where preserving natural language diversity is critical. By reducing reliance on English structure, DAS-generated dialogues may serve as a valuable resource for improving multilingual dialogue systems, enabling models to better capture the linguistic diversity needed for effective cross-lingual communication.

G Prompts

G.1 Paraphrase

Produce a new conversation from the given dialogue by paraphrasing each utterance.

Conversation:
<conversation>

G.2 Machine Translation

Translate the following conversation into <language>.

Conversation:

<conversation>

G.3 Single Prompt Localize+Translate

Translate the following conversation into <language>. While translating, please localize the dialogue for <language> speakers. This should include any necessary changes to names, locations, social dynamics, common objects (replace any brands or items with more commonly used ones), and general cultural appropriateness to make the context feel natural for <language> speakers. Assign culturally appropriate names based on gender, age, and relationship dynamics in the target culture. Be mindful of specifying politeness levels, family dynamics, and relevant cultural norms.
Conversation:
<conversation>

G.4 Encode

You will read dialogue snippets. Assign a function label to each utterance with all necessary parameters to reconstruct the meaning. The goal is to capture what the speaker is doing (e.g., asking a question, making a request, giving feedback) rather than how they say it. The 'parameters' of the functions will be whatever is necessary to capture the meaning of the utterance. This should be the minimum amount of information necessary to convey all of the information of the sentence.

Here is the complete list of functions with descriptions and examples:

<function name>: <description>
- example: <example>
...

Note: It's possible for one utterance (or even one sentence) to serve multiple purposes. In this case, it's fine to choose more than one, but keep them in the order presented.
Example:
text: "No, I don't think so",
functions: ["disagree()", "express(doubt)"]

Conversation:
<conversation>

G.5 Generate Context

Summarize the scene by creating details about the characters to capture the context of the dialogue. If a name is provided, use that, but if not, feel free to make up details. Don't use the same names as the example. Provide at minimum, each speaker's name, gender (M,F,X), age, and presumed relationship to the other speaker. Try to capture the context of the scene. Don't let every conversation be between a man and a woman. Try to vary up the gender combinations.

Example:
Two coworkers, Alex (M, 35) and Jamie (X, 28), are discussing a project deadline and planning next steps. Alex is a project manager, Jamie is a software developer. The conversation takes place in the office break room, where they often chat about after-work activities.

Conversation:
<conversation>

G.6 Localize Context

You will be provided with a scenario in which a dialogue is taking place. Please localize the dialogue context for <language> speakers. This should include any necessary changes to names, locations, social dynamics, common objects (replace any brands or items with more commonly used ones), and general cultural relevance to make the context feel natural for <language> speakers. Assign culturally appropriate names based on gender, age, and relationship dynamics in the target culture. Be mindful of specifying politeness levels, family dynamics, and relevant cultural norms.
Do NOT write a sample conversation. Only provide the localized scenario.

Scenario:
<context>

Target language/culture: <language>

G.7 Localize DAS

Please localize the following Dialogue Act Script for <language> speakers while maintaining the original structure and meaning. Do not remove, condense, or add new topics. Only adjust cultural

references when necessary, and keep all turns intact. The format must remain exactly the same, with only localized modifications where relevant.

Target language/culture: <language>
Summary: <localized context>

DAS:
<DAS turns>

G.8 Decode

You are given a conversation setting with details about the speakers, their ages, genders, and relationships. Use this information to generate the text of the conversation based on the provided functions for each turn. Consider the speakers' ages, relationships, and any relevant details to make the conversation natural and contextually accurate. It is okay to leave out or make up parts of the functions if they don't fit what the characters would naturally say. Aim for cultural authenticity even if the names of the characters/places/foods need to be changed.

You don't have to stick to one function per sentence. Some functions will combine naturally into a single sentence.
Example:
functions: A.disagree(); A.express(doubt)
A: "No, I don't think so"

Do not merge multiple turns into a single response. Maintain the same turn structure. Ensure that each turn corresponds to an individual line of dialogue. Do not repeat or shorten any of the functions or dialogue history.

Language: <language>
Context: <localized context>
Conversation:
<localized DAS turns>