Improved Dialogue Localization and Translation 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, which can introduce artifacts that reduce their naturalness and cultural relevance. This work proposes Dialogue Act Script (DAS) as a structured framework for encoding, localizing, and generating multilingual dialogues. Rather than directly translating, DAS generates new dialogues in the target language by adapting a language-independent representation, ensuring greater cultural relevance and naturalness. By using structured dialogue act representations, DAS improves multilingual dialogue localization by enhancing cultural adaptability, reducing translationese, and providing an interpretable framework for structured adaptation. The results show that DAS-generated dialogues consistently outperform machine and human translations across Italian, German, and Chinese in all evaluation criteria, particularly in cultural relevance, coherence, and situational appropriateness, suggesting that functional abstraction allows explicit adaptation to conversational norms that straightforward machine translation may not capture.[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 approach is translating existing English dialogue datasets, but direct translation 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 translation artifacts that make dialogues sound unnatural in the target language (Artetxe et al., 2020).

For instance, translated dialogues may still be set in American locations, reference brands unfamiliar to speakers of the target language, or use names that are common in English-speaking countries but rare elsewhere, leaving the dataset culturally English-speaking even after translation. This limits its usefulness for training and evaluating dialogue systems in diverse linguistic and cultural settings. Additionally, Majewska et al. (2023) found that English-to-Russian translations often retained passive voice constructions that are typical in English but unnatural in spoken Russian, making the dialogues sound stiff and formal.

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

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

questions:

1. How does encoding dialogues with DAS influence the quality of the generated dialogue?
2. To what extent does DAS improve the interpretability and control of multilingual dialogue generation?
3. What are the trade-offs between structured and flexible function labeling in DAS, and how do they impact reproducibility and dialogue quality?
4. How well can automated evaluation methods leveraging LLMs, approximate human judgments of dialogue quality?

By addressing these questions, we aim to demonstrate that DAS improves multilingual dialogue localization through both automated and human evaluations across multiple languages. To evaluate our approach, we use XDailyDialog (Liu et al., 2023), which provides professional translations of DailyDialog (Li et al., 2017) in Italian, German, and Chinese.

Our experiments assess how well the original content is retained in the encoding process and evaluate the quality of DAS-generated dialogues in the target language based on cultural relevance, fluency, situational appropriateness, and coherence. Our results show that DAS-generated dialogues consistently outperform both machine and human translations across all evaluation criteria, particularly in cultural relevance, coherence, and situational appropriateness. This demonstrates that DAS's structured abstraction enables explicit adaptation to conversational norms, overcoming the limitations of direct translation and ensuring more natural, contextually appropriate dialogue generation.

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.

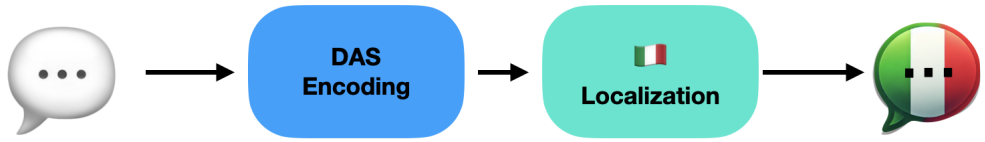
Recent work has explored how LLMs can generate structured representations from natural language. Li et al. (2023) turned information extraction into a code generation task, using Code-LLMs to produce structured outputs. Similarly, Sainz et al. (2024) introduced GoLLIE, a guideline-aware LLM for zero-shot 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. DAS represents communicative intent using a predefined set of dialogue acts and parameters. Dialogue acts categorize utterances based on their communicative function—such as requesting, informing, or directing—rather than their surface form (Austin, 1962).

Although DAS has potential applications in dialogue planning and grammatical error correction, this work focuses on its use for multilingual dialogue localization and translation. DAS enables culturally adaptive dialogue generation, reducing reliance on direct translation, which often fails to capture pragmatic and sociocultural nuances. By abstracting dialogue into structured representations, DAS helps mitigate anglocentric biases in multilingual datasets, improving translation naturalness and cross-linguistic authenticity.



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

3.2 Dialogue Translation with DAS

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 meaning while allowing for flexible reconstruction across languages and cultural contexts. For example, the English utterance "Then I'll just use my credit card." may be encoded as `inform(subject=self, action=use_credit_card, condition=insufficient_cash, timeframe=future)`.

Localization: The DAS representation is then adapted to align with cultural norms of the target language by modifying specific parameters, such as named entities, cultural references, or commonly used objects, while preserving the original dialogue act and intent. For instance, when adapting for a Chinese audience, the action `use_credit_card` might change to `use_Alipay_or_WeChat_Pay`, reflecting more commonly used payment methods in China.

Decoding: The final step generates fully realized dialogue in the target language based on the localized DAS representation. This process reconstructs the conversation in a way that is fluent, coherent, and contextually appropriate, while maintaining alignment with the original communicative intent. For example, the localized representation `inform(subject=self, action=use_Alipay_or_WeChat_Pay, condition=insufficient_cash, timeframe=future)` would be decoded into Chinese as: 我会用支付宝或者微信支付来付款。 (I will use Alipay or WeChat Pay to make the payment.)

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 classified based on its communicative function (e.g., inquire, express, agree), which then determines the corresponding function in the script. This ensures that the speaker's intent remains intact across different phrasings and linguistic realizations.

While numerous dialogue act taxonomies have been established, including CUED Standard Dialogue Acts (Young, 2009), DIT++ Taxonomy (Bunt et al., 2020), and the Schema-Guided Dialogue (SGD) dataset (Shah et al., 2018), our study instead

evaluates how well GPT can annotate dialogues using an unseen, task-specific set of dialogue acts. This approach allowed us to assess its adaptability to a newly defined schema rather than measuring performance against established classification standards. The schema was developed iteratively through human-in-the-loop refinement (Monarch, 2021): initial dialogue act categories were generated by prompting ChatGPT with example conversations, followed by a pilot human annotation phase. Categories with low inter-annotator agreement (e.g., `explain` was found to be difficult to distinguish from `inform` or `clarify`) were removed, and annotators were given the option to propose new dialogue acts when none of the existing ones fit. This process ensured that the final schema balanced flexibility with consistency while remaining informed by real conversational data. For the full list of 15 dialogue acts in our annotation schema and corresponding examples, see Appendix A.

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—such as entities, actions, and contextual references—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 E.

3.4 Localization

Localization encourages cultural adaptability, producing less direct translations that enable the creation of multilingual datasets that avoid the anglocentric biases that typically result from direct translations from English. However, DAS also allows for more direct translations while maintaining natural phrasing, offering flexibility depending on the intended use case.

For localization, we tested GPT-4o and GPT-4o-mini (OpenAI et al., 2024) by prompting the LLM to first localize the context by making necessary adjustments to names, locations, social dynamics, and commonly referenced objects, such as replacing brands or items with ones more familiar in the target culture. Additionally, the LLM is instructed to ensure general cultural relevance, making the context feel natural in the target language. The LLM also localizes the DAS turns by updating the parameters—such as replacing `location=New York` with `location=Beijing`—while keeping the dialogue acts themselves unchanged.

3.5 Decoding

Decoding involves generating the target language dialogue from the DAS representation. Given the character descriptions and setting, which may have been localized, the LLM generates a possible utterance for each turn of the conversation. This process allows for additional constraints to be applied, such as adjusting the difficulty of the language to suit specific needs. For example, the DAS encoding: `inquire(topic=menu, subject=house_specials)` could be realized in different ways: with simple grammar and vocabulary ("Do you have house specials?"), or a more polite, complex version ("Would you be able to tell me about the house specials currently on offer?") By controlling the level of complexity, DAS can generate dialogues that match the needs of language learners or different conversational contexts.

Decoding can be conducted turn by turn, for instance, if the dialogue is ongoing and a single DAS turn is generated as part of a chatbot's response. Alternatively, the entire dialogue can be decoded at once for localization purposes. Dialogues can be generated in any language supported by the LLM. For this study, we tested GPT-4o and GPT-4o-mini for Chinese, Italian, German, and En-

glish, including Simple English to evaluate whether DAS effectively supports language simplification.

4 Experiments

4.1 Experimental Setup

For our evaluation, 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 compare DAS localization against professional human translations provided by XDailyDialog (Liu et al., 2023) in Italian, German, and Chinese. Additionally, we include a simple machine translation baseline, generated by passing the dialogue to GPT-4o and prompting it to translate into the target language. Prompts can be found in Appendix G.2.

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.

4.2 Encoding Consistency

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 Closed DAS function labeling. We used Cohen's Kappa (κ) (Cohen, 1960) as the evaluation metric, which accounts for chance agreement in categorical annotations.

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.

Two human annotators labeled 105 dialogue turns from five randomly selected conversations, using a predefined set of DAS functions. Both annotators were provided with the full set of DAS function definitions and illustrative examples to ensure consistent understanding. The same definitions and examples were provided to GPT-4o and GPT-4o-mini, ensuring that humans and models followed identical annotation guidelines. The results are shown in Table 1.

The high agreement between human annotators (κ = 0.844) in Closed DAS suggests that a structured function set ensures annotation consistency, making it a viable framework for reliable dialogue encoding. Human-GPT agreement in Closed DAS remains substantial (κ = 0.822 with Human1, κ = 0.769 with Human2), confirming that LLMs can effectively apply predefined DAS categories when provided with clear definitions and examples.

When comparing human-GPT agreement, GPT-4o achieved a higher alignment than GPT-4o-mini, suggesting that more capable LLMs better capture DAS functions when explicitly prompted. However, even GPT-4o-mini maintains substantial agreement (κ = 0.765, 0.746). The results of this experiment support the use of our chosen set of functions for dialogue act annotation.

4.3 Decoding Back into English

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

[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 $50 USD on experiments.

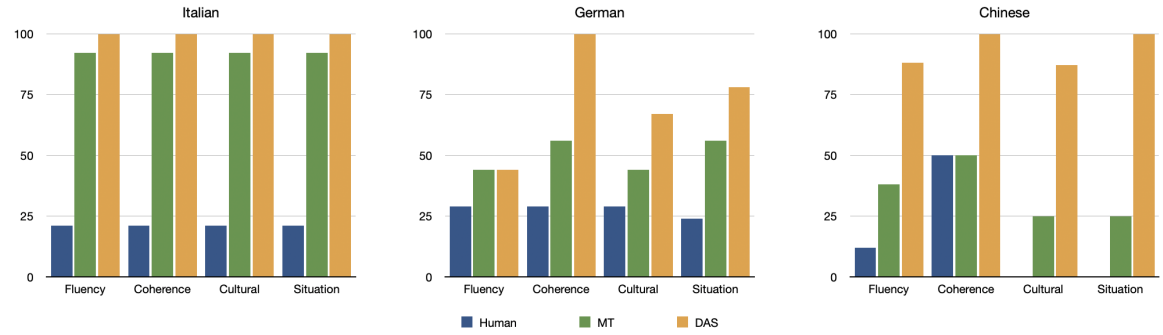


Figure 2: Win rates of each system across evaluation criteria (fluency, coherence, cultural relevance, and situational appropriateness). Higher win rates indicate stronger performance in pairwise comparisons.

Metric	DAS	Original
Fluency	0.727	0.455
Coherence	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.

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 "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 C for details and results of this experiment.

4.4 Localization Quality

To assess the quality of localized and translated dialogues produced by our DAS-based method, we conducted a human evaluation using Amazon Mechanical Turk (MTurk)[3] and recruited annotators. Native speakers of Chinese, Italian, and German were asked to compare DAS-localized dialogues against direct translations, evaluating each conversation's fluency, coherence, cultural relevance, and situational appropriateness.

Annotators were presented with a random pair of translations from the DailyDialog dataset (8-16 turns per conversation) and asked the following questions[4]:

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

[3] https://www.mturk.com

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

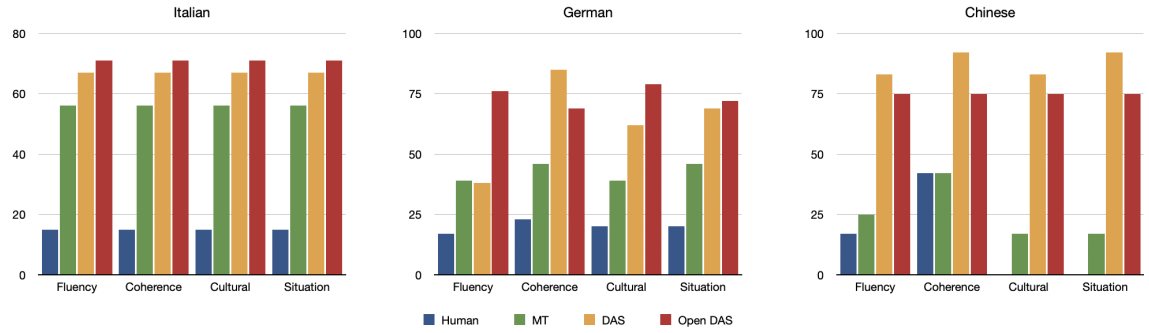


Figure 3: Win rates of each system, including Open DAS, across evaluation criteria.

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 presented with randomized conversation pairs and allowed to select A, B, Both, or Neither for each question. Win rates are calculated as in section 4.3, with "Both" counted as a win for both systems, and "Neither" counting as a loss for both.

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. Unlike direct translation, DAS does not reference the original wording during decoding, allowing for greater flexibility in how dialogues are realized. This enables shifts in style and expression that better align with the conversational norms of the target language, rather than being constrained by the source-language phrasing.

4.5 Open DAS

The Closed DAS framework enforces a fixed set of dialogue functions, ensuring high reproducibility but potentially limiting expressiveness. While this structured approach benefits annotation consistency and automation, it may overconstrain how communicative intent is represented.

To explore whether a more flexible annotation scheme could capture richer dialogue dynamics, we introduce Open DAS—a variant of DAS where the model defines dialogue acts freely rather than selecting from a predefined set. While this allows for greater expressiveness, it may reduce annotation consistency and reproducibility. To quantify this variability, we computed Cohen's Kappa between the human-annotated Closed DAS function labels and GPT-generated Open DAS encodings. Table 3 reports the findings:

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

Table 3: 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.

As expected, the results show a sharp decline in annotation agreement when using Open DAS. An inspection of the data revealed that this is likely because the model has a tendency to include extra information in the function name (e.g., `offer_assistance` instead of `offer`). When function labels were truncated by keeping only the part before an underscore, agreement improved to κ = 0.269, suggesting that at least some of the disagreement stemmed from the model introducing fine-grained distinctions between dialogue acts.

Figure 3 reports human preferences for Open DAS, Closed DAS, machine translation, and professional human translation across the same four evaluation criteria outlined in Section 4.4. Preference for Open DAS varied by language: in German and Italian, Open DAS slightly outperformed Closed DAS in all criteria. However, in Chinese, Open DAS was ranked lower than Closed DAS in all categories.

Method	Cosine Similarity	KL-Divergence
Human	0.825	0.014
MT	0.912	0.006
DAS	0.650	0.030

Table 4: Statistical analysis of cosine similarity and KL-Divergence between English source texts and their Italian translations from XDailyDialog. All values statistically significant ($p < 0.0001$).

4.6 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 4 presents the results of the analysis of Italian data. German and Chinese yielded similar results that can be found in the Appendix F.

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.

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.

5 Conclusion

This study introduced Dialogue Act Script as a structured approach to dialogue abstraction and explored its application to dialogue localization. DAS-based translations consistently outperformed standard MT in cultural relevance, coherence, and situational appropriateness, suggesting that functional abstraction allows for explicit adaptation to conversational norms that straightforward translation may not capture.

A key advantage of DAS is its modularity and reusability. Unlike direct translation, which must be performed separately for each language pair, DAS encoding occurs only once and can be adapted to multiple target languages, making it a cost-effective and scalable alternative for multilingual applications.

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 exist within the scope of our work. One such limitation is the inability to fully verify annotation quality in crowd-based assessments, particularly for Chinese. For Italian and German, we restricted MTurk participation to workers located in the respective countries, but this was not possible for Chinese due to platform availability. As a result, we could not control the geographical location of Chinese annotators, making it difficult to verify annotation quality. The interannotator agreement (Krippendorff's Alpha) was so low for Chinese that we removed the crowdworker data from the experiment and recruited an annotator that we could confirm was a native Chinese speaker. While Italian and German used multiple annotators for each conversation pair, Chinese was limited to a single annotation.

Another limitation concerns the computational cost of DAS decoding, which currently relies on LLMs to generate output. While DAS encoding is reusable across languages, deploying DAS in low-resource settings remains challenging due to the dependence on high-quality generative models. Exploring lighter-weight generation strategies could improve accessibility in multilingual applications.

This study evaluated DAS using a single dataset (XDailyDialog), which consists of chitchat-style dialogues. While this dataset is useful for conversational settings, DAS's applicability to other domains remains untested. Future work should assess whether DAS encoding and localization strategies generalize to task-oriented dialogues, such as customer service, medical, or legal interactions, where conversational constraints may differ.

While DAS enables cultural adaptation, its approach to localization has not been extensively evaluated for potential biases in cultural representation. Ensuring that localized dialogues align with cultural norms without reinforcing stereotypes remains an open challenge. Furthermore, DAS has primarily been tested on well-resourced languages, and its effectiveness for low-resource or morphologically complex languages remains uncertain. Future work should examine how well DAS encoding generalizes to languages with fewer training resources or different structural properties.

Ethical Considerations

Our study involved human annotations for evaluating DAS-generated dialogues. We recruited crowdworkers via Amazon Mechanical Turk (MTurk) for Italian, German, and Chinese evaluations.

Crowdworkers were compensated $1.20 USD per conversation pair, with an estimated 5 minutes of work per task. To further validate results, we hired one native speaker each for German, Chinese, and English, along with a contributing author who participated in English evaluations. These annotators were voluntary participants, compensated at $12 USD per hour.

Our study adhered to ethical guidelines for fair compensation and informed consent. Workers participated voluntarily and were informed of the nature of the task, with no foreseeable risks of harm.

While no explicitly harmful outputs were observed, LLM-generated text presents inherent risks of unintended biases, particularly in speaker roles and cultural adaptations. One notable pattern was a strong tendency for the LLM to assume one speaker was male and the other female, leading to skewed conversational distributions. Despite mitigation efforts, this bias persisted, with 88% of conversations featuring male-female pairings.

Additionally, while cultural adaptations were designed to align with local norms, we have not exhaustively searched for potential biases or harmful stereotypes in localized dialogues. As LLMs reflect biases present in their training data, future work should further investigate these risks. We caution potential adopters of this framework to critically examine LLM outputs for unintended biases and take proactive measures to ensure fair and accurate representations across languages and cultures.

While this work focuses on improving multilingual dialogue generation, we acknowledge potential risks related to bias and misuse. Future work should explore bias mitigation strategies and safeguards against potential misuse.

Finally, as DAS relies on large-scale LLMs for encoding, localization, and decoding, its computational demands contribute to the broader environmental concerns associated with energy-intensive NLP models. Future research could explore lighter-weight models or efficiency optimizations to make multilingual dialogue adaptation more sustainable.

The XDailyDialog dataset is used under the Apache-2.0 License, which permits research and commercial use with proper attribution. Our use of the dataset for evaluating multilingual dialogue adaptation aligns with its intended purpose as a resource for dialogue system research.

The DailyDialog dataset (which XDailyDialog

is build on) is licensed under CC BY-NC-SA 4.0. The original copyright of all English conversations belongs to the source owner. The dataset consists of crawled conversations from websites designed to help English learners practice conversational English through roleplay. It primarily contains chitchat-style dialogues and may not represent a diverse range of conversational domains.

While the DAS framework enables cultural adaptation of dialogues, it is not intended for high-stakes applications where misinterpretations of localized meaning could have real-world consequences, such as legal, medical, or financial translations. Any future deployment outside research contexts should include additional safeguards and human validation to ensure responsible use.

AI assistance from ChatGPT and GitHub Copilot was used for minor language adjustments in writing and line-level code completion. However, all research ideas, code architecture, and experimental design were solely the author's work, and all AI-assisted outputs were thoroughly vetted for correctness.

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.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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.

Robert Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning Publications.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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.

Steve Young. 2009. Cued standard dialogue acts. Technical report, Cambridge University Engineering Department.

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 Human Evaluation

To improve data reliability, we implemented the following participation restrictions:

- Workers were required to have at least a 95% approval rating and a minimum of 100 completed tasks.
- For Italian and German, workers were limited to users in Italy and Germany, respectively.
- Since MTurk is unavailable in China, we did not enforce geographic restrictions for Chinese evaluations but required self-reported native fluency.

Despite these precautions, inter-annotator agreement (Krippendorff's Alpha) among crowdworkers alone varied significantly across languages. While Italian showed moderate agreement (0.413), Chinese (0.066) and German (0.114) were near-random, indicating inconsistencies in how annotators judged translation quality.

For German, we identified one MTurk worker who selected "Both" for every question, suggesting a lack of engagement with the task. This worker was excluded from the analysis and replaced by a new annotator, after which Krippendorff's Alpha increased from 0.114 to 0.815, reflecting a substantial improvement in annotation reliability.

For Chinese, each conversation was annotated by a different crowdworker, preventing direct inter-annotator agreement comparisons. Due to this single-annotator-per-sample setup, we were unable to assess annotation consistency or verify quality. As a result, we removed the MTurk annotations for Chinese entirely and relied only on the expert annotator for evaluation.

C 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 temper-

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

ature 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 5.

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.

D 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 6.

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 6: 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.

The results indicate strong alignment between 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.

E 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 7: Win rates against machine translation and human translation for including a context summary or not.

The results in Table 7 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 D).

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

F Multilanguage Experiments

Method	Cos Sim.	KL Div.
Italian (as shown in Table 4)		
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 8: 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 ($p < 0.0001$).

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 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.4 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.5 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.6 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.7 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>