Cross-Lingual Dialogue Dataset Creation via Outline-Based Generation

Olga Majewska[◇] **Evgeniia Razumovskaia**[◇] **Edoardo M. Ponti**[†◇]
Ivan Vulić[◇] **Anna Korhonen**[◇]
[◇]Language Technology Lab, University of Cambridge, United Kingdom
[†]Institute for Language, Cognition and Computation, University of Edinburgh, United Kingdom
{om304,er563,ep490,iv250,alk23}@cam.ac.uk

Abstract

Multilingual task-oriented dialogue (ToD) facilitates access to services and information for many (communities of) speakers. Nevertheless, its potential is not fully realized, as current multilingual ToD datasets—both for modular and end-to-end modeling—suffer from severe limitations. **1)** When created from scratch, they are usually small in scale and fail to cover many possible dialogue flows. **2)** Translation-based ToD datasets might lack naturalness and cultural specificity in the target language. In this work, to tackle these limitations we propose a novel *outline-based* annotation process for multilingual ToD datasets, where domain-specific abstract schemata of dialogue are mapped into natural language outlines. These in turn guide the target language annotators in writing dialogues by providing instructions about each turn's intents and slots. Through this process we annotate a new large-scale dataset for evaluation of multilingual and cross-lingual ToD systems. Our **C**ross-lingual **O**utline-based **D**ialogue dataset (COD) enables natural language understanding, dialogue state tracking, and end-to-end dialogue evaluation in 4 diverse languages: Arabic, Indonesian, Russian, and Kiswahili. Qualitative and quantitative analyses of COD versus an equivalent translation-based dataset demonstrate improvements in data quality, unlocked by the outline-based approach. Finally, we benchmark a series of state-of-the-art systems for cross-lingual ToD, setting reference scores for future work and demonstrating that COD prevents over-inflated performance, typically met with prior translation-based ToD datasets.

1 Introduction and Motivation

One of the staples of machine intelligence is the ability to communicate with humans and complete a task as instructed during such an interaction. This is commonly referred to as task-oriented dialogue (ToD; Gupta et al., 2005; Bohus and Rudnicky 2009; Young et al., 2013; Muise et al., 2019). Despite having far-reaching applications, such as banking (Altinok, 2018), travel (Zang et al., 2020), and healthcare (Denecke et al., 2019), this technology is currently accessible to very few communities of speakers (Razumovskaia et al., 2022a).

The progress in multilingual ToD is critically hampered by the paucity of training data for many of the world's languages. While cross-lingual transfer learning (Zhang et al., 2019; Xu et al., 2020; Siddhant et al., 2020; Krishnan et al., 2021) offers a partial remedy, its success is tenuous beyond typologically similar languages and generally hard to assess due to the lack of evaluation benchmarks (Razumovskaia et al., 2022a). What is more, transfer learning often cannot leverage multi-source transfer and few-shot learning due to lack of language diversity in the ToD datasets (Zhu et al., 2020; Quan et al., 2020; Farajian et al., 2020).

Therefore, the main driver of development in multilingual ToD is the creation of multilingual resources. However, even when available, they suffer from several pitfalls. Most are obtained by manual or semi-automatic translation of an English source (Castellucci et al., 2019; Bellomaria et al., 2019; Susanto and Lu, 2017; Upadhyay et al., 2018; Xu et al., 2020; Ding et al., 2022; Zuo et al., 2021, *inter alia*). While this process is cost-efficient and typically makes data and results comparable across languages, it yields dialogues that lack *naturalness* (Lembersky et al., 2012; Volansky et al., 2015), are not properly *localized* nor *culture-specific* (Clark et al., 2020). Further, they provide over-optimistic estimates of performance due to the artificial similarity between source and target texts (Artetxe et al., 2020). As an alternative to translation, new ToD datasets can be created from scratch in a target language through the Wizard-of-Oz framework (WOZ; Kelley

Transactions of the Association for Computational Linguistics, vol. 11, pp. 139–156, 2023. https://doi.org/10.1162/tacl_a_00539
Action Editor: Hai Zhao. Submission batch: 3/2022; Revision batch: 9/2022; Published 1/2023.

Outlines	Dialogue & Slot Output
USER: *Express the desire to search for roundtrip flights for a trip*	Мне нужно найти рейс в Ставрополь и обратно авиакомпании S7.
the name of the airport or city to arrive at: Seattle the company that provides air transport services: American Airlines	Ставрополь S7
ASSISTANT/SYSTEM: *Inform the user that you found 1 such option(s). Offer the following option(s):*	Найден 1 рейс авиакомпании S7 с пересадкой, вылет в 7:35, возвращение в Москву в 16:15. Стоимость билетов 6845 рублей.
the company that provides air transport services: American Airlines departure time of the flight flying to the destination: 7:35am departure time of the flight coming back from the trip: 4:15pm the total cost of the flight tickets: $343	S7 07:35 16:15 6845 рублей

Table 1: Example from the COD dataset of outline-based dialogue generation in Russian with target language substitutions of slot values. The first column (**Outline**) includes example outlines presented to the dialogue creators, and the second column holds the creators' output (**Dialogue & Slot Output**).

1984) where humans impersonate both the client and the assistant. However, this process is highly *time- and money-consuming*, thus *failing to scale* to large quantities of examples and languages, and often *lacks coverage* in terms of possible dialogue flows (Zhu et al., 2020; Quan et al., 2020).

To address all these gaps, in this work we devise a novel *outline-based* annotation pipeline for multilingual ToD datasets that combines the best of both processes. In particular, abstract *dialogue schemata*, specific to individual domains, are sampled from the English Schema-Guided Dialogue dataset (SGD; Shah et al., 2018; Rastogi et al., 2020). Then, the schemata are automatically mapped into outlines in English, which describe the intention that should underlie each dialogue turn and the slots of information it should contain, as shown in Table 1. Finally, outlines are paraphrased by human subjects into their native tongue and slot values are adapted to the target culture and geography. This ensures both the cost-effectiveness and cross-lingual comparability offered by manual translation, and the naturalness and culture-specificity of creating data from scratch. Through this process, we create the **C**ross-lingual **O**utline-based **D**ialogue dataset (termed COD), supporting natural language understanding (intent detection and slot labeling tasks), dialogue state tracking, and end-to-end dialogue modeling in 11 domains and 4 typologically and areally diverse languages: Arabic, Indonesian, Russian, and Kiswahili.

To confirm the advantages of the leveraged annotation process, we run a proof-of-concept experiment where we create two analogous datasets through the outline-based pipeline and manual translation, respectively. Based on a quality survey from human participants, we find that, while having similar annotation speed, outline-based annotation achieves significantly higher naturalness and familiarity of concepts and entities, without compromising data quality and language fluency.[1] Finally, crucial evidence showed that cross-lingual transfer test scores on translation-based data are over-estimated. We demonstrate that this is due to the fact that the distribution of the sentences (and their hidden representations) is considerably more divergent between training and evaluation dialogues in COD than in the translation-based dataset.

Further, to establish realistic estimates of performance on multilingual ToD, we benchmark a series of state-of-the-art multilingual ToD models in different ToD tasks on COD. Among other findings, we report that zero-shot transfer surpasses 'translate-test' on slot labeling, but this trend is reversed for intent detection. Language-specific performance also varies substantially among evaluated models, depending on the quantity of unlabeled data available for pretraining.

In sum, COD provides a typologically diverse dataset for end-to-end dialogue modeling and evaluation, and streamlines a scalable annotation process that results in natural and localised dialogues. We hope that COD will contribute to democratizing dialogue technology and facilitating reliable cost-effective ToD systems for a wide array of languages. Our data and code are available at `github.com/cambridgeltl/COD`.

[1]Furthermore, when asked to compare equivalent dialogues obtained with the two processes, respondents favored outline-based dialogues in more than 80% cases.

2 Related Work

Although a number of NLU resources have recently emerged in languages other than English, the availability of high-quality, multi-domain data to support multilingual ToD is still inconsistent (Razumovskaia et al., 2022a). Translation of English data has been the predominant method for generating examples in other languages: For example, the ATIS corpus (Hemphill et al., 1990) boasts translations into Chinese (He et al., 2013), Vietnamese (Dao et al., 2021), Spanish, German, Indonesian, and Turkish, among others (Susanto and Lu, 2017; Upadhyay et al., 2018; Xu et al., 2020). Bottom-up collection of ToD data directly in the target language has been the less popular choice (e.g., in French [Bonneau-Maynard et al., 2005] and Chinese [Zhang et al., 2017; Gong et al., 2019]).

Concurrent work by FitzGerald et al. (2022) employs translation as part of a dataset creation workflow where Amazon MTurk workers first translate or localize slot values, and subsequently translate or localize entire phrases in which these slots appear. While localization allows improving the geographical and cultural relevance of entities mentioned in dialogues, this approach still relies on translation from English, thus perpetuating many of the problems of earlier translation-based methods: For example, introducing English grammatical and lexical biases in dialogue utterances (Koppel and Ordan, 2011) or compromising target language idiomacity. As we demonstrate in §4, our outline-based dialogue generation method addresses these issues by eschewing direct translation in favor of guided dialogue creation in the target language, ensuring naturalness of linguistic expressions used in each language and yielding a dataset better capturing linguistic diversity.

Thus far, the focus of existing benchmarks has been predominantly either on monolingual multi-domain (Hakkani-Tür et al., 2016; Liu et al., 2019; Larson et al., 2019) *or* multilingual single-domain evaluation (Xu et al., 2020), rather than balancing diversity along both these dimensions. Moreover, the current multilingual datasets are mostly constrained to the two NLU tasks of intent detection and slot labeling (Li et al., 2021; van der Goot et al., 2021), and do not enable evaluations of E2E ToD systems in multilingual setups. In order to adequately assess the strengths and generalizability of NLU as well as DST and E2E models, they should be tested both on multiple languages *and* multiple domains, a goal pursued in this work.

3 Annotation Design

We selected the English Schema-Guided Dialogue (SGD) dataset (Shah et al., 2018; Rastogi et al., 2020) as a starting point due to its scale (20k human-assistant dialogues) and diversity (20 domains). It was constructed via automatic generation of *dialogue schemata* combined with manual creation of dialogue paraphrases by crowdworkers, organized as lists of turns for each individual interaction, each turn containing an utterance by the user or system. The accompanying annotations are grouped into frames, each corresponding to a single API or service (e.g., *Banks_2*). In turn, each service is represented as a schema including its characteristic functions (intents) and parameters (slots), as well as their natural language (NL) descriptions.[2]

We first assessed the viability of our method on Russian, collecting data using (i) direct translation from English and (ii) our proposed outline-based approach. We then applied our method to three other languages that boast a large number of speakers and yet suffer from a shortage of resources: Arabic, Indonesian, and Kiswahili, ensuring the dataset's diversity in terms of language family and macro-area, as well as writing systems (Cyrillic, Arabic, and Latin scripts), see Table 2.[3] In Table 3 we quantify the linguistic diversity of the language sample and compare it with the standard multilingual dialogue NLU and end-to-end datasets. In terms of typology, COD is comparable to datasets with much larger language samples (e.g., Multi-ATIS++, xSID) and considerably exceeds others. With respect to family and macroarea diversity, COD is the most diverse out of existing datasets.

3.1 Data Creation Protocol

The data creation protocol involved the following phases: **1)** source dialogue sampling, **2)** automatic generation of outlines based on intent and slot information using rewrite rules, **3)** manual outline-driven target language dialogue creation

[2]For example, the ''*Alarm_1*'' service comprises intents such as ''*GetAlarms*'' (''*Get the alarms user has already set*'') and ''*AddAlarm*'' (''*Set a new alarm*'') and slots ''*alarm_time*'', ''*alarm_name*'', ''*new_alarm_time*'', and ''*new_alarm_name*''.

[3]The total cost of COD was 800 GBP per language.

Language	ISO	**Family**	**Branch**	**Macro-area**	**L1** [M]	**Total** [M]
Russian	RU	Indo-European	Balto-Slavic	Eurasia	153.7	258
Standard Arabic	AR	Afro-Asiatic	Semitic	Eurasia / Africa	0†	274
Indonesian	ID	Austronesian	Malayo-Polynesian	Papunesia	43.6	199
Kiswahili	SW	Niger–Congo	Bantu	Africa	16.3	69

Table 2: Language statistics. The last two columns denote the number of speakers in millions. †Standard Arabic is learned as L2.

	NLU-Only Datasets					**End-to-End Datasets**		
	M. TOP	M. ATIS	MultiATIS++	MTOP	xSID	BiTOD	GlobalWOZ	COD
# languages	3	3	9	6	13	2	3	4
Typology	0.20	0.29	0.33	0.29	0.37	0.15	0.24	0.31
Family	0.67	0.67	0.44	0.33	0.50	1.0	0.75	1.0
Macroareas	0	0	0	0	0.26	0	0.14	1.04

Table 3: Comparison of diversity indices of multilingual dialogue datasets in terms of typology, family, and macroareas. For the description of the three diversity measures, we refer the reader to Ponti et al. (2020). M. TOP was created by Schuster et al. (2019); M. ATIS (Upadhyay et al., 2018); MultiATIS++ (Xu et al., 2020); MTOP (Li et al., 2021); xSID (van der Goot et al., 2021); BiTOD (Lin et al., 2021); GlobalWOZ (Ding et al., 2022).

	Alarm (♢)	Flights	Homes	Movies	Music	Media	Banks	Payment (♢)	RideSharing	Travel	Weather	**#turns**
Dev	13	12	12	16	14	–	14	–	–	12	18	**1138**
Test	21	23	13	19	16	17	–	8	11	–	–	**1352**

Table 4: Number of dialogues per domain and total number of turns in each set. ♢ marks the domains that are not included in the (English) training set.

and slot annotation, and **4)** post-hoc review, all described here.

Source Dialogue Sampling. To ensure wide coverage of dialogue scenarios, we randomly sampled source dialogues from across 11 domains, out of which five (*Alarm, Flights, Homes, Movies, Music*) are shared between the development and test set; the remainder are unique to either set, to enable *cross-domain* experiments. To guarantee a balanced coverage of different intents, we sampled 10 examples per intent, which ensures the task cannot be solved by simply predicting the most common intent (see Table 4 for dataset statistics).

Outline Generation. Our goal was to create minimal but sufficient instructions for target language dialogue creators to ensure coverage of specific intents and slots, while avoiding imposing predefined syntactic structures or linguistic expressions. First, for each user or system act, we manually created a rewrite rule, for example, INFORM_COUNT→*Inform the user that you found* + `INFORM_COUNT[value]` + *such option(s)* (`value` corresponds to the number of options matching the user request). Next, we automatically match each intent and slot with its NL description (provided in the SGD schemata) and used them to generate intent/slot-specific outlines (with stylistic adaptations where necessary): For example, an intent ''SearchOnewayFlight'' and a description ''*Search for one-way flights to the destination of choice*'' would yield an outline *Express the desire to search for one-way flights* (see Table 5).

Dialogue Writing. We recruited target language native speakers fluent in English via the `proz.com` platform.[4] Dialogue creators were presented with language-specific guidelines.[5] An essential part of the task consisted in a cultural adaptation of culturally and geographically

[4]To ensure quality, we selected candidates with reported target language credentials who successfully completed a qualification exercise consisting in writing a 6-turn dialogue according to outlines analogous to those in the main task.

[5]`github.com/evgeniiaraz/Supplementary`.

Act	Slot/Intent	Description	Value	Outline
INFORM_INTENT	SearchOnewayFlight	Search for one-way flights to the destination of choice	– –	***Express the desire to*** *search for one-way flights*
REQUEST	number_checked_bags	Number of bags to check in	2	***Ask if*** *the number of bags to check in is 2*

Table 5: Examples of dialogue generation outlines created from SGD schemata, that is, annotations of dialogue acts, intents, slots and values, with intent-specific rewrites in bold.

specific slot values (e.g., city names, movie titles) through substitutions with named entities more familiar or closer to the creators' culture (e.g., American Airlines→Aeroflot, New York→Jakarta).

Slot Span Validation. First, creators performed slot span labeling while working on dialogue writing. Subsequently, the annotated data in each language underwent an additional round of manual revision by a target language native speaker and a final automatic check for slot value-span matches. We verified inter-annotator reliability on Russian, where we collected slot span annotations from pairs of independent native-speaker annotators. The accuracy scores (i.e., ratio of slot instances with matching spans to the total annotated instances) of 0.99 for development data and 0.98 for test data reveal very high agreement on this task.

4 Translation versus Outline-Based

The main motivation behind the outline-based approach is to avoid the known pitfalls of *direct translation* and produce evaluation data better representing the linguistic and cultural realities of each language in the sample. To verify whether the method satisfies these goals in practice, we carried out a trial experiment consisting in parallel dialogue data creation using two different methods, (i) direct translation and (ii) outline-based generation, starting from the same sample of source SGD dialogues to ensure a fair comparison. In (i), randomly sampled (see §3.1) English user/system utterances were extracted directly from the SGD data with accompanying slot and intent annotations and subsequently translated into the target language by professional translators, also responsible for validating target language slot spans. In (ii), we automatically extracted dialogue frames, including intents and slots, matching dialogue IDs sampled in (i), and used them to generate NL outlines to guide manual dialogue creation by native speakers (§3.1).

Questions
Q1. The ASSISTANT helps satisfy the USER's requests.
Q2. The USER speaks naturally and sounds like a Russian native speaker.
Q3. The ASSISTANT speaks naturally and sounds like a Russian native speaker.
Q4. I can easily imagine myself mentioning or hearing the proper names referred to in the dialogue (e.g., titles of films or songs, people, places) in a conversation with my Russian friends or family.

Table 6: Quality survey questions (Part 1).

We also asked the participants to time themselves while working on the task. Notably, we found the annotation speed to be identical for the two methods, averaging 15 seconds per single dialogue turn (dialogue writing + slot annotation). While the translation approach does not require any creative input in terms of cultural adaptations of slot values, the outline-based approach allows freedom in terms of the linguistic expressions used, which results in similar time requirements.

Quality Survey. We assessed the quality of the two methods' output in a survey with 15 Russian native speakers, consisting of (1) independent and (2) comparative evaluation.[6] Within each part, the order of questions was randomized. In Part 1, the respondents were presented with 6 randomly sampled dialogues from the data generated by either method (3 dialogues per method) and asked to answer to what extent they agree with each of four statements in Table 6 (translated into Russian) by giving a 1-5 rating. In Part 2, respondents were presented with 5 randomly sampled pairs of matching dialogue excerpts from both datasets (based on

[6]The non-comparative part came first to avoid priming effects from an a priori awareness of systematic qualitative differences between examples coming from either method.

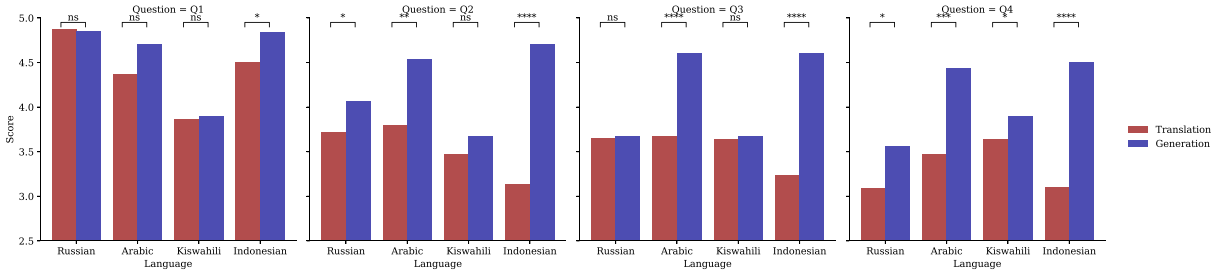


Figure 1: Average scores for each quality survey question (see Table 6) assigned to dialogue examples generated via *translation* versus *outline*-based generation in each language. Statistically significant differences (paired Student's *t*-test) are indicated as follows: $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), $p \leq 0.0001$ (****); *ns* indicates $p > 0.05$.

shared dialogue IDs) and asked to choose which excerpt (A or B) sounded more natural to them. Following the validation experiments and analyses of our outline-based method in Russian (as reported in the remainder of §4), we extended the quality survey to the other three languages included in COD, Arabic, Kiswahili, and Indonesian, comparing outline-generated dialogues to those translated from English by professional translators in an analogous two-part evaluation setup. All survey questions and instructions were translated into each target language and 15 native speaker participants were recruited for each language-specific survey.

Figure 1 shows average scores for Part 1 questions (Q1–Q4) across the 15 participants in each language. The methods produce dialogues which score similarly in terms of the assistant's goal-orientedness (Q1), with a statistically significant negative effect of translation, with respect to outline-based generation, noted only in Indonesian. However, we observe consistent differences in the perceived naturalness and target-language fluency (Q2 and Q3). First, the user utterances created based on outlines are perceived as more natural-sounding (Q2) across all four languages, with the largest quality gap observed in Indonesian and Arabic. This pattern is repeated for Q3, where Arabic and Indonesian participants found outline-based generated assistant utterances substantially closer to natural target language spoken by native speakers than their translated counterparts.

Crucially, outline-generated dialogues score consistently better in terms of the familiarity of mentioned entities (Q4), with significant score differences found in all four languages. These results are encouraging, given that Q4 directly addresses one of the main objectives of our method, namely, target language-specificity. While both approaches are capable of producing convincing dialogues in each language, as reflected in positive (>3) average scores, it is worth noting that the perceived degree of naturalness and familiarity of the conversations is on average lower in the case of Kiswahili. This emphasizes the need for careful debiasing of the concepts and situation types referred to in the dialogues, to ensure that the entire dialogues scenarios, not just slot values, reflect the linguistic and cultural reality of target language communities.

The patterns noticed in the independent evaluation (Part 1) are further reinforced in the results of the comparative evaluation in Part 2, even more clearly skewed in favor of the outline-based method. Out of 75 comparisons (15 participants judging 5 pairs each) in each language, outline-based dialogues are judged as more natural-sounding, on average, in over 80% of cases, with a near-perfect preference found in Indonesian (94%), followed by Arabic (82%), Russian (80%), and Kiswahili (76%). Table 7 shows an example pair of matching dialogue excerpts from each method with accompanying English translations.

Effects of *Translationese*. Dialogue data should be representative of natural interactions between two interlocutors. The utterances of both the user *and* the system should reflect the properties characteristic of the conversational register in a given language, appropriate for the communicative situation at hand and the participants' social roles (Chaves et al., 2019; Chaves and Gerosa, 2021). When qualitatively comparing the translation and outline-based generation in Table 7, we observe that translated utterances are often skewed to the source language syntax and lexicon (known as

Translation	Outline-based Generation
USER: Можете ли вы найти мне квартиру, в которой можно держать домашних животных? ♣	USER: Мне нужно найти квартиру в аренду в которой можно селиться с животными.
EN: *Can you find me an apartment that allows pets?*	*I need to find a flat to rent in which I can live with pets.*
ASSISTANT: Alma Apartments предлагают 1 спальню с 1 ванной комнатой за 1 900$ в месяц. <...>	ASSISTANT: Позволь мне предложить следующий вариант. Это однокомнатная квартира с одной ванной комнатой. Квартира находится в жилом комплексе Белая вишня. <...>
EN: *Alma Apartments has a 1 bedroom 1 bathroom for 1,900 per month. <...>*	*Let me propose the following option. It is a 1 bedroom 1 bathroom apartment. The apartment is located in the "White Cherry" building. <...>*
USER: Звучит многообещающе. Я бы хотел взглянуть на неё. Не могли бы вы назначить визит на 10-е число? <...> Если подумать, то лучше 12-го числа этого месяца.	USER: Отлично. Можно договориться о просмотре квартиры 10-го числа? <...> Ой не, лучше на 12 число.
EN: *Sounds promising. I'd like to look it over. Could you schedule a visit on the 10th? <...> On second thought, make it the 12th of this month.*	*Perfect. Could we arrange the visit on the 10th? <...> Oh no, better on the 12th.*
ASSISTANT: Итак, вы хотите посетить Alma Apartments 12 марта.	ASSISTANT: Резервирую осмотр квартиры в жилом комплексе Белая вишня на 12 марта?
EN: *Okay, you would like to visit Alma Apartments on March 12th.*	*Am I booking the visit to "White Cherry" on the 12th of March?*
USER: Да. Какой там номер телефона?	USER: Да, именно так. Можешь найти контактный номер телефона?
EN: *Yes. What is their phone number?*	*Yes, exactly. Could you find their contact number?*
ASSISTANT: 650-813-1369. Ваш визит запланирован. ♣♠	ASSISTANT: Да, номер телефона 650-813-1369. Я успешно забронировала осмотр.
EN: *It's 650-813-1369. Your visit is scheduled.*	*Yes, the phone number is 650-813-1369. I successfully booked the visit.*

Table 7: Comparison of dialogues generated by each method. For each user/assistant utterance, we provide the original English sentences from SGD for the translation method, and English translations of the Russian utterances written based on outlines. ♣ – syntactic similarity to source language; ♠ – lexical similarity to source language.

the ''translationese'' effects [Koppel and Ordan, 2011]), compromising fluency and idiomacity that are essential in natural-sounding exchanges.

One issue which arises in literal translation is syntactic calques from the source language. For instance, the translation of the first USER utterance (Table 7, col. 'Translation') uses a dative pronoun найти мне [DATIVE] (*find me*), even though the transitive verb найти (*find*) does not require the [DATIVE] case after it—a likely calque of the English expression *Can you find me*. In comparison, the corresponding outline-based generated utterance uses a more natural construction. Another problem concerns the differences in the use of grammatical structures depending on the language register. For instance, using passive voice in spoken English is common: For example, the last ASSISTANT utterance in Table 7. Its translation into Russian also includes passive voice, although it is usually avoided in spoken Russian (Babby and Brecht, 1975). In contrast, the outline-based utterance uses a simpler active voice construction, preserving the original meaning.

Lexical ''translationese'' effects include (i) the preference for lexical cognates of source language words, and (ii) the use of a vocabulary typical for the written language, both exemplified by the last ASSISTANT utterance (Table 7). The translation includes the verb запланирован (*is planned*), even though the verb планировать, having the same root as English *to plan*, is rarely used in spoken Russian when arranging near-future appointments and more frequently when making a step-by-step plan. In contrast, the outline-based generated utterance includes the verb забронировать (*to book*) which is more specific to arranging appointments and more frequently used in spoken language.

Slot Localization. Datasets collected via translation stay largely grounded in the realm of the Anglosphere (Zuo et al., 2021; Hung et al., 2022). For instance, slot values are directly translated rather than being substituted with a culture-specific equivalent. As a result, multilingual models are tested in a very favorable context where only the surface language changes but the entities stay the same (this bias is especially pertinent for models in cross-lingual setups). In COD guidelines, annotators are explicitly instructed to replace English concepts with their target language equivalents. In this study, we calculate the percentage of slot values which were localised. We consider a slot value to be localized if the

Split	Localized Slot Values				
	AR	ID	RU	SW	AVG
Dev	42.98	59.68	61.60	76.51	60.19
Test	13.50	57.00	53.81	78.34	50.66

Table 8: Per-language percentage of localised slot values in the COD dataset.

value is conceptually different from its English counterpart (e.g., using a local artist's name or converting a sum in GBP or USD to the local currency). Table 8 demonstrates that more than half of all slot values in the dataset are localized, which is a large improvement. This shows that with the COD dataset models will be tested on more culturally and linguistically aware data than if the dataset were created via translation.

Evaluation of Translation-Based vs. Outline-Generated Data. The vast majority of existing NLU datasets are based on translation from English to the target language (Xu et al., 2020; van der Goot et al., 2021). This could lead to an overly optimistic evaluation of cross-lingual ToD systems, since the data might not be representative of real-life language use, due to ''translationese'' effects discussed above. We verify this hypothesis in the following diagnostic experiment. We use a *translate-train* approach where: (i) training data are translated from the source language (`en`) to the target (`ru`) via Google Translate; and (ii) the model is fine-tuned on these automatically translated data. We then test the model on evaluation data obtained by: **(a)** translation using Google Translate, **(b)** translation by professional translators (closest in nature to existing dialogue NLU datasets), **(c)** generated based on outlines. For the experiment, we fine-tune mBERT (Devlin et al., 2019) on intent detection.[7]

The results in Table 9 show a stronger performance on translation-based evaluation sets than on more natural, outline-based generated examples, thus corroborating previous observations in other areas of NLP, e.g., machine translation (Graham et al., 2020), now also attested in ToD. Crucially, this experiment verifies that using solely translation-based ToD evaluation data might lead to an inflated estimation of models' cross-lingual capabilities and, consequently, too optimistic performance expectations in real-life applications. This further validates our proposed outline-based approach to multilingual ToD data creation.

Data Creation	Split	Accuracy
Google Translate	Dev	47.98
	Test	35.06
Professional Translation	Dev	48.33
	Test	34.62
Outline-based Generation	Dev	40.25
	Test	31.81

Table 9: Cross-lingual intent detection accuracy on development and test data **(a)** translated via Google Translate; **(b)** translated by professionals; and **(c)** outline-generated: COD.

Analysis of Sentence Encodings. One reason behind the scores in Table 9 likely lies in the differences between multilingual sentence encodings of English examples, examples generated via translation, and those yielded by the outline-based method. To test this, we obtain sentence encodings of all user turns for one intent from the three datasets via the distilled multilingual USE sentence encoder (Yang et al., 2020; Reimers and Gurevych 2019).[8]

As shown in Figure 2, as expected, the translation-based data are encoded into sentence representations that are much more similar to their English source than the corresponding outline-generated examples. We use pairwise KL-divergence scores between KDE-estimated Gaussians to measure the similarity between English (En), Translated to Russian (Trans), and Outline-based sentences: KL (En || Trans) $= 7.5 \times 10^{-4}$; KL (En || Outline) = 4.69×10^{-5}; KL (Trans || Outline) $= 3.84 \times 10^{-5}$. As expected, direct translation artificially skews target utterances towards English. This again reinforces the finding from Table 9: Multilingual ToD datasets collected via outline-based generation should lead to more realistic assessments of multilingual ToD models than their translation-based counterparts.

[7]We focus on the intent detection task to avoid the interference of noise introduced by the alignment algorithms (i.e., aligning the source language examples with automatic translations of the training data for slot labeling).

[8]The same trends were observed in the results with other standard multilingual sentence encoders such as LaBSE (Feng et al., 2022), not included due to space limits.

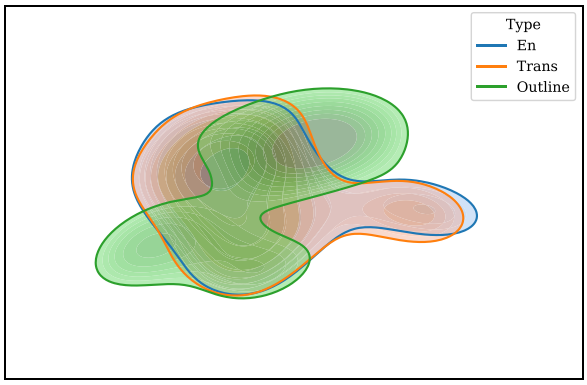


Figure 2: Kernel density estimate (KDE) plot for distributions of user turn encodings via the distilled multilingual USE. Input sentences are either the original sentences in English (`En`), translated to Russian (`Trans`), or generated in Russian based on Outlines (`Outline`). Dimensionality reduction was performed using tSNE (Van der Maaten and Hinton, 2012).

5 Baselines, Results, Discussion

COD includes labeled data for three standard ToD tasks: i) Natural Language Understanding (NLU; intent detection and slot labeling); ii) dialogue state tracking (DST); and iii) end-to-end (E2E) dialogue modeling. Here, we benchmark a representative selection of state-of-the-art models (§5.1) on our new dataset, highlighting its potential for evaluation and the key challenges it presents across different tasks and experimental setups (§5.2).

Notation. A dialogue $\mathcal{D}$ is a sequence of alternating user and system turns $\{\mathcal{U}_1, \mathcal{S}_1, \mathcal{U}_2, \mathcal{S}_2, ...\}$. Dialogue history at turn t is the set of turns up to point t, i.e., $\mathcal{H}_t = \{\mathcal{U}_1, \mathcal{S}_1, \ldots, \mathcal{U}_{t-1}, \mathcal{S}_{t-1}, \mathcal{U}_t\}$.

5.1 Baselines and Experimental Setup

We evaluate and compare the baselines for each task along the following axes: (i) different multilingual pretrained models; (ii) cross-lingual transfer approaches; (iii) in-domain versus cross-domain.

Multilingual Pretrained Models. For cross-lingual transfer based on multilingual pretrained models, we abide by the standard procedure where the entire set of encoder parameters and the task-specific classifier head are fine-tuned. We evaluate the following pretrained language models: (i) for NLU and DST, we use the Base variants of multilingual BERT (mBERT; Devlin et al., 2019) and XLM on RoBERTa (XLM-R; Conneau et al., 2020); the models were pretrained on Wikipedia in over 100 languages and CommonCrawl dataset, respectively; for intent detection and slot labeling, we evaluate both a model that jointly learns the two tasks (Xu et al., 2020) as well as separate task-specific models; (ii) for E2E modeling, we use multilingual T5 (mT5; Xue et al., 2021), a sequence-to-sequence model demonstrated to be the strongest baseline for cross-lingual dialogue generation (Lin et al., 2021).

Cross-lingual Transfer. We focus on two standard methods of cross-lingual transfer: (i) transfer based on multilingual pretrained models and (ii) *translate-test* (Hu et al., 2020). In (i), a Transformer-based encoder is pretrained on multiple languages with a language modeling objective, yielding strong cross-lingual representations that enable zero-shot model transfer. In (ii), test data in a target language are translated into English via a translation system: We compare Google Translate (GTr)[9] and MarianMT (Junczys-Dowmunt et al., 2018). The models in both transfer methods are fine-tuned on the original English task-specific data from the English SGD dataset.

For end-to-end training, we set up two additional cross-lingual baselines, similar to Lin et al. (2021). In few-shot fine-tuning (FF), after the model is trained on source language data (EN), it is further fine-tuned on a small number of target language dialogues. In our FF experiments, we use the dev sets in each language as few-shot learning data. In mixed-language pretraining (MLT; Lin et al., 2021), the model is fine-tuned on mixed language data where the slot values in the source language data are substituted with their target language counterparts. Unlike Lin et al. (2021), we do not assume the existence of a bilingual parallel knowledge base, unrealistic for low-resource languages. Hence, the translations of slot values are obtained via MarianMT (Junczys-Dowmunt et al., 2018).

In-Domain versus Cross-Domain Experiments. COD development and test splits include examples belonging to domains which were not seen in the English training data (see Table 4). This enables cross-lingual evaluation in 3 different regimes: *in-domain* testing (**In**), where the model

[9] `cloud.google.com/translate/docs/apis`.

Setup	TrSystem	Model	Intent Detection *(Accuracy)* AR	ID	RU	SW	AVG	Slot Labeling (F_1) AR	ID	RU	SW	AVG
MEncoder		mBERT	18.61	17.57	22.83	6.09	16.28	21.54	15.29	24.89	8.84	17.64
TrTest	GTr	mBERT	24.46	27.34	28.97	23.93	26.18	11.70	16.36	19.56	16.67	16.07
	MarianMT	mBERT	28.40	26.89	29.38	25.38	27.51	13.28	14.89	20.21	11.98	15.09
MEncoder		XLM-R	25.56	29.88	27.60	19.59	25.66	28.65	31.73	32.47	15.18	27.00
TrTest	GTr	XLM-R	27.43	29.53	29.76	26.42	28.29	10.61	19.55	18.70	14.94	15.95
	MarianMT	XLM-R	29.20	29.11	30.53	26.39	28.81	13.10	16.96	18.35	11.27	14.92

Table 10: Per-language NLU results for (i) zero-shot cross-lingual transfer using multilingual pretrained models (MEncoder) and (ii) translate-test (TrTest) transfer with Google Translate and MarianMT (see §5.1). Translations for slot labeling were aligned using *fast_align* (Dyer et al., 2013). MEncoder results are from the *separate* training regime (see §5.1). All scores are averages over 5 random seeds and follow the **All**-domain setup.

is evaluated on examples coming from the domains seen during training; *cross-domain testing* (**Cross**), evaluating on examples coming from the domains which were *not* seen during training; and *overall* testing (**All**), evaluating on all examples in the evaluation set.

Architectures and Training Hyperparameters. NLU in ToD consists of two tasks performed for each user turn $\mathcal{U}_i$: intent detection and slot labeling, which are typically framed as sentence- and token-level classification tasks, respectively. When a model is trained in a joint fashion, the two tasks share an encoder, and task-specific classification layers are added on top of the encoder (Zhang et al., 2019; Xu et al., 2020). The loss is a sum of the intent classification and the slot labeling losses (cross-entropy). In *separate* training, there is no parameter sharing, so neither NLU task influences the other. The performance metrics are *accuracy* for intent detection and F_1 for slot labeling.

In the DST task, the model maps the dialogue history $\mathcal{H}_t$ to the belief state at $\mathcal{U}_t$; this includes the slot values that have been filled up to turn t. We use BERT-DST (Chao and Lane, 2019) in the experiments, which makes a binary classification regarding the relevance of every slot-value pair to the current context. During training, negative dialogue context-slot pairs are sampled randomly in a 1:1 ratio. At inference time, every context is mapped to every possible slot-value pair. The performance metric used for DST is the standard Joint Goal Accuracy (JGA) (Rastogi et al., 2020), defined as the ratio of dialogue turns in which all slot values are correctly predicted.

As in prior work (Lin et al., 2021), E2E modeling is framed as a sequence-to-sequence (seq2seq) generation task. At every turn t, the goal is to predict the following $\mathcal{S}_t$ based on $\mathcal{H}_t$ fed into the model as a concatenated string. We adopt the generative seq2seq model, termed mSeq2Seq, as used by Lin et al. (2021). This is based on mT5 Small and mT5 Base (Xue et al., 2021) and standard top-k sampling. Unless stated otherwise, Small version of the model is used. As in prior work (Lin et al., 2021), performance is reported as BLEU scores (Papineni et al., 2002). Unless stated otherwise, we use a beam size of 5 for generation.[10]

Source Language Training. We train all models on the standard full training split of the English SGD dataset (Rastogi et al., 2020). In order to measure performance gaps due to transfer and ensure comparability of dialogue flows in all languages, we also evaluate on the subset of the English SGD test set sampled as a source for COD (see Table 4).

5.2 Results and Discussion

Below we discuss the results of cross-lingual transfer under the experimental setups in §5.1. We report both per-language scores and averages across the four COD target languages.

Main Results. Table 10 compares the results for the two NLU tasks, while Table 11 shows scores

[10] We opt for mT5 as it substantially outperformed mBART (Liu et al., 2020a) and other E2E baselines in the work of Lin et al. (2021). We leave experimentation with more sophisticated model variants (Liu et al., 2020b) and sampling methods such as nucleus sampling (Holtzman et al., 2020) for future work. For brevity, we do not report results with other automatic E2E modeling metrics such as Task Success Rate or Dialogue Success Rate (Budzianowski and Vulić, 2019).

Setup	Model	E2E Training *(BLEU)* AR	ID	RU	SW	AVG
MEncoder	mT5	0.90	2.06	1.63	1.79	1.60
+FF	mT5	4.36	10.96	8.48	7.79	7.90
+MLT +FF	mT5	4.26	10.40	9.00	7.02	7.67
TrTest GTr	mT5	1.87	1.96	4.38	2.59	2.59
MarianMT	mT5	1.74	1.74	4.08	1.42	2.25

Table 11: Per-language E2E results for two cross-lingual transfer methods (see also the information in Table 10).

	Dialog State Tracking *(JGA)* AR	ID	RU	SW	AVG
mBERT	0.44	0.00	0.01	0.17	0.16

Table 12: Baseline results for DST on COD test set using mBERT as an encoder.

in the E2E task. With translate-test (TrTest), the gains are highly task-dependent: It performs considerably better than encoder-based (MEncoder) transfer on intent detection and E2E modeling, while the opposite holds for slot labeling. This is likely because: **1)** we rely on a word alignment algorithm on top of English predictions to align them with the target language, which adds noise to the final predictions; and **2)** many errors are due to incorrect 'label granularity' (e.g., predicting *departure city* instead of *departure airport*), as shown by qualitative analysis.[11] Note that TrTest, unlike MEncoder, assumes access to high-quality MT systems and/or parallel data for different language pairs.

Table 11 reveals large gains of TrTest over the vanilla version of MEncoder, both with MarianMT and GTr, but GTr proves consistently better: This corroborates recent findings on other cross-lingual NLP tasks (Ponti et al., 2021). However, the +FF results in Table 11 reverse this trend and underline the benefits of few-shot target language fine-tuning in E2E training. The performance gains are large, even though the target language data include only 92 dialogues (<1% of English training data). In contrast, +MLT does not have a significant impact, possibly due to i) noisy target language substitutes, obtained via automatic translation, unlike in Lin et al. (2021) where ground truth target language slot values were available; or ii) culture-specificity of slot values in COD. Thus, substitution with translations seems beneficial only for dialogues with a pre-defined common cross-lingual slot ontology.

Figure 3c presents another interesting trend, concerning the comparison of E2E performance of a larger versus a smaller model: mT5-Base versus mT5-Small. While zero-shot performance is comparable between the two, we observe that mT5-Base performs considerably better in a few-shot training scenario (+FF). We hypothesize that in zero-shot training the models overfit to generation in English,[12] while in few-shot training the model's cross-lingual generation capabilities are highlighted, once the model has encountered several examples in the target language.

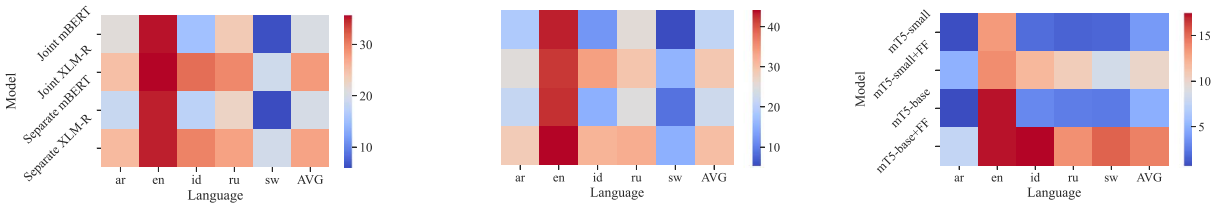
In DST, irrespective of the transfer method and target language, cross-lingual performance is near-zero, as visible from Table 12. These findings are in line with prior work (Ding et al., 2022) and are due to the DST task complexity. This is even more pronounced in zero-shot cross-lingual settings and especially for COD, where culture-specific slot values are obtained via outline-based generation. Given the low results, we focus on NLU and E2E as the two main tasks in all the following analyses.

Comparison of Multilingual Models on NLU. The results in Table 10 and Figure 3 indicate that XLM-R largely outperforms mBERT in all setups in both NLU tasks, especially on two languages more distant from English, ID and SW. We attribute this to XLM-R being exposed to more data in these languages during pretraining than mBERT. This very reason also accounts for the discrepancy in their performance on EN relative to other languages: With XLM-R, the gap between EN scores and other languages is much smaller than with mBERT. This is especially apparent in the case of Indonesian: ID pretraining data for mBERT are less than 10% of EN pretraining data, while their sizes are comparable in XLM-R.

Further, the results in Figure 3 indicate that joint training of two NLU tasks tends to benefit intent detection while degrading the performance on slot labeling. The reverse trend is true for separate training: Slot labeling scores improve, while intent detection degrades. This confirms

[11] This is more likely in translated text where language-specific hints for the exact slot type may get lost in translation.

[12] The observation is also corroborated by weaker performance in languages which use non-Latin script.



(a) Intent Detection (b) Slot Labeling (c) E2E

Figure 3: Per-language results over all domains. (a) and (b) share the model labels on the *y*-axis.

the trend observed in recent work (Razumovskaia et al., 2022b).[13]

Gaps with Respect to English. The per-language NLU results (Table 10 and Figure 3) also illustrate a performance gap due to 'information loss' during transfer: The drops (averaged across all 4 target languages) of the strongest transfer method are ≈10 points on intent detection (in **All**-domains experiments), and 15 points on slot labeling, using exactly the same underlying models. These gaps are even more pronounced for some languages (e.g., the lowest-resource language Kiswahili) and in domain-specific setups (e.g., **In**-domain setups).

The E2E results in Figure 3c also reveal a chasm between mT5 performance on English and the other four languages, especially so without any target-language adaptation. The gap, while still present, is substantially reduced with the +FF model variant (see §5.1). This disparity emphasizes the key importance of (i) continuous development of multilingual benchmarks inclusive of less-resourced languages to provide realistic estimates of performance on multilingual ToD, as well as (ii) creation of (indispensable) in-domain data for few-shot target language adaptation. The low absolute scores indicate the complexity of the task in general. Overall, these findings reveal the challenging nature of COD, and call for further research on data-efficient and effective transfer methods in multilingual ToD.

In-Domain vs. Cross-Domain Evaluation. COD not only enables cross-lingual transfer but is also the first multilingual dialogue dataset suitable for testing models in cross-domain settings

[13]We also evaluated whether incorporating English SGD schemata into the NLU models—that is, leveraging short English descriptions of domains, intents, and slots available from the English SGD dataset—improves performance, adapting the process of Cao and Zhang (2021) to a cross-lingual setup; however, we obtained negative results.

Method	Model	In	Cross	All
Intent Detection (*Accuracy*)				
Joint	mBERT	15.85	13.19	15.22
	XLM-R	28.45	20.55	26.30
Separate	mBERT	17.28	12.50	15.08
	XLM-R	28.97	19.51	25.66
Slot Labeling (F_1)				
Joint	mBERT	18.80	15.21	15.60
	XLM-R	29.85	26.13	26.80
Separate	mBERT	21.75	16.73	17.64
	XLM-R	30.22	26.31	27.01
E2E (*BLEU*)				
	mT5-small	1.40	1.47	1.60
	mT5-base	2.06	2.20	2.29

Table 13: Baseline results for NLU and E2E on the COD test set, averaged over all 4 target languages; **In**-, **Cross**-domain, and **All** domains setups.

(Table 13). The general observation is that in-domain performance is much higher than cross-domain, although both have large room for improvement.

We conduct a more detailed analysis of the in-domain and cross-domain performance for the slot labeling task. We chose to focus on slot labeling as the annotators were explicitly instructed to substitute slot values with target language-specific values where appropriate. We use XLM-R fine-tuned on the full English dataset. In the interest of space and clarity we present the results for two domains that the model has seen in training (`Flights`, `Movies`) and one domain which it has *not* seen during training (`Payment`). The results in Table 14 support the general claims: There is a significant drop between domains seen and not seen at training.[14] Further, we note that the performance on `Flights` is much lower than on `Movies`. This is due to: (i) the larger number of

[14]The results on `Payment` are lower than the averaged Cross-domain scores in Table 13, as the test set for Table 13 also included examples not assigned to any domain, which were assigned the `NONE` label.

Domain	Slot Labeling (F_1) AR	ID	RU	SW	AVG
Flights	26.67	20.00	24.71	7.45	19.70
Movies	18.97	39.25	29.06	26.89	28.54
Payment	3.20	5.12	2.42	0.70	2.86

Table 14: Results for cross-lingual slot labeling for 3 domains: Flights, Movies (**In**-domain) and Payment (**Cross**-domain).

slots in the `Flights` domain; (ii) the slot values in `Flights` are naturally suited for localization (e.g., departure and destination cities) which makes the domain more complex for cross-cultural generalisation. This additionally proves the need to collect multilingual dialogue datasets in a more culturally aware fashion to get realistic estimates of cross-lingual performance of ToD models.

6 Conclusion and Outlook

We have presented and validated a 'bottom-up' method for the creation of multilingual task-oriented dialogue (ToD) datasets. The key idea is to map domain-specific language-independent dialogue schemata to natural language outlines, which in turn guide human dialogue generators to create natural target-language utterances, for the user and system alike. We have empirically demonstrated that the proposed outline-based approach yields more natural and culturally sensitive dialogues than the standard translation-based approach to multilingual ToD data creation. Moreover, we have proven that the standard translation-based approaches often yield over-inflated and unrealistic performance in multilingual evaluation, while this issue is removed with the outline-based generation method.

Our proposed approach yielded a new **C**ross-lingual **O**utline-based **D**ialogue dataset (termed COD), which covers 5 typologically diverse languages, 11 domains in total, and enables evaluations in standard NLU, DST, and end-to-end ToD tasks. Thus, COD is an important step towards challenging multilingual *and* multi-domain ToD evaluation in future research. We have also evaluated a series of state-of-the-art models for the different ToD tasks, setting baseline reference points, and revealing the challenging nature of the dataset with ample room for improvement.

We hope that our work will inspire future research across multiple aspects. One such area concerns cultural debiasing of the concepts and situations captured in the dialogues. Our method addresses this through cultural adaptations and replacements of foreign concepts with those common in the annotators' culture and environment. The next step should involve a careful selection of dialogue scenarios based on their relevance and plausibility in the culture in question, as very recently started in other NLP areas (e.g., Liu et al., 2021). In this work, we presented useful practices and insights hoping to guide similar (potentially larger-scale) data creation efforts in ToD for other, especially lower-resource, languages, and domains.

COD is available online at `github.com/cambridgeltl/COD`.

Acknowledgments

This work was funded by the ERC PoC Grant MultiConvAI: Enabling Multilingual Conversational AI (no. 957356) and a research donation from Huawei. The work of EMP was supported by the Facebook CIFAR AI Chair program. We would like to thank our annotators for their contribution to this work and the TACL editors and anonymous reviewers for their helpful feedback and suggestions.

References

Duygu Altinok. 2018. An ontology-based dialogue management system for banking and finance dialogue systems. In *Proceedings of the First Financial Narrative Processing Workshop (FNP)*, pages 1–9.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of EMNLP 2020*, pages 7674–7684. `https://doi.org/10.18653/v1/2020.emnlp-main.618`

Leonard H. Babby and Richard D. Brecht. 1975. The syntax of voice in Russian. *Language*, pages 342–367. `https://doi.org/10.2307/412860`

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019.

Almawave-SLU: A new dataset for SLU in Italian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*.

Dan Bohus and Alexander I. Rudnicky. 2009. The RavenClaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23(3):332–361. https://doi.org/10.1016/j.csl.2008.10.001

Helene Bonneau-Maynard, Sophie Rosset, Christelle Ayache, A. Kuhn, and Djamel Mostefa. 2005. Semantic annotation of the French Media Dialog Corpus. In *Proceedings of the Eurospeech 2005*, pages 3457–3460. https://doi.org/10.21437/Interspeech.2005-312

Paweł Budzianowski and Ivan Vulić. 2019. Hello, It's GPT-2-How can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22. https://doi.org/10.18653/v1/D19-5602

Jie Cao and Yi Zhang. 2021. A comparative study on schema-guided dialogue state tracking. In *Proceedings of NAACL-HLT 2021*, pages 782–796. https://doi.org/10.18653/v1/2021.naacl-main.62

Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multi-lingual intent detection and slot filling in a joint BERT-based model. *arXiv preprint arXiv:1907.02884*.

Guan-Lin Chao and Ian Lane. 2019. BERT-DST: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. In *Proceedings of Interspeech 2019*, pages 1468–1472. https://doi.org/10.21437/Interspeech.2019-1355

Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's how you say it: Identifying appropriate register for chatbot language design. In *Proceedings of HAI'19*, HAI '19, pages 102–109. https://doi.org/10.1145/3349537.3351901

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. Why should we care about register? Reflections on chatbot language design. *arXiv preprint arXiv:2104.14699*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *TACL*, 8:454–470. https://doi.org/10.1162/tacl_a_00317

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent detection and slot filling for Vietnamese. In *Proceedings of Interspeech 2021*, pages 4698–4702.

Kerstin Denecke, Mauro Tschanz, Tim Lucas Dorner, and Richard May. 2019. Intelligent conversational agents in healthcare: Hype or hope? *Studies in Health Technology and Informatics*, 259:77–84.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, volume 1, pages 4171–4186.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of ACL 2022*, pages 1639–1657. https://doi.org/10.18653/v1/2022.acl-long.115

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of NAACL-HLT 2013*, pages 644–648.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022.

Language-agnostic BERT sentence embedding. In *Proceedings of ACL 2022*, pages 878–891.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2022. MASSIVE: A 1M-Example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*.

Yu Gong, Xusheng Luo, Yu Zhu, Wenwu Ou, Zhao Li, Muhua Zhu, Kenny Q. Zhu, Lu Duan, and Xi Chen. 2019. Deep cascade multi-task learning for slot filling in Chinese E-commerce Shopping Guide Assistant. In *Proceedings of AAAI 2019*, volume 33, pages 6465–6472. https://doi.org/10.1609/aaai.v33i01.33016465

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of NAACL-HLT 2021*, pages 2479–2497. https://doi.org/10.18653/v1/2021.naacl-main.197

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of EMNLP 2020*, pages 72–81. https://doi.org/10.18653/v1/2020.emnlp-main.6

Narendra Gupta, Gokhan Tur, Dilek Hakkani-Tur, Srinivas Bangalore, Giuseppe Riccardi, and Mazin Gilbert. 2005. The AT&T spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):213–222. https://doi.org/10.1109/TSA.2005.854085

Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM. In *Proceedings of Interspeech 2016*, pages 715–719. https://doi.org/10.21437/Interspeech.2016-402

Xiaodong He, Li Deng, Dilek Hakkani-Tur, and Gokhan Tur. 2013. Multi-style adaptive training for robust cross-lingual spoken language understanding. In *Proceedings of ICASSP 2013*, pages 8342–8346. IEEE.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24–27, 1990*, pages 96–101. https://doi.org/10.3115/116580.116613

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of ICLR 2020*, pages 1–16.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML 2020*, pages 4411–4421.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Ponzetto, and Goran Glavaš. 2022. Multi2WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In *Proceedings of NAACL-HLT 2022*, pages 3687–3703.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121. https://doi.org/10.18653/v1/P18-4020

John F. Kelley. 1984. An Iterative Design Methodology for User-friendly Natural Language Office Information Applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41. https://doi.org/10.1145/357417.357420

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the ACL-HLT 2011*, pages 1318–1326.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223. https://doi.org/10.18653/v1/2021.mrl-1.18

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1311–1316. https://doi.org/10.18653/v1/D19-1131

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825. https://doi.org/10.1162/COLI_a_00111

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of EACL 2021*, pages 2950–2962.

Zhaojiang Lin, Andrea Madotto, Genta Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale N. Fung. 2021. BiToD: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of EMNLP 2021*, pages 10467–10485.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of IWSDS 2019*, pages 165–183.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *TACL*, 8:726–742.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020b. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of AAAI 2020*, pages 8433–8440. https://doi.org/10.1609/aaai.v34i05.6362

Laurens Van der Maaten and Geoffrey Hinton. 2012. Visualizing non-metric similarities in multiple maps. *Machine Learning*, 87(1):33–55. https://doi.org/10.1007/s10994-011-5273-4

Christian Muise, Tathagata Chakraborti, Shubham Agarwal, Ondrej Bajgar, Arunima Chaudhary, Luis A. Lastras-Montano, Josef Ondrej, Miroslav Vodolan, and Charlie Wiecha. 2019. Planning for goal-oriented dialogue systems. *arXiv preprint arXiv:1910.08137*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318. https://doi.org/10.3115/1073083.1073135

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of EMNLP 2020*, pages 2362–2376.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modeling latent translations for cross-lingual transfer. *arXiv preprint arXiv:2107.11353*.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. RiSAWOZ: A large-scale multi-domain wizard-of-oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of EMNLP 2020*, pages 930–940. https://doi.org/10.18653/v1/2020.emnlp-main.67

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of AAAI 2020*,

pages 8689–8696. https://doi.org/10.1609/aaai.v34i05.6394

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo M. Ponti, Anna Korhonen, and Ivan Vulic. 2022a. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *Journal of Artificial Intelligence Research*, 74:1351–1402. https://doi.org/10.1613/jair.1.13083

Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2022b. Data augmentation and learned layer aggregation for improved multilingual language understanding in dialogue. In *Findings of ACL 2022*, pages 2017–2033. https://doi.org/10.18653/v1/2022.findings-acl.160

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3982–3992. https://doi.org/10.18653/v1/D19-1410

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of NAACL-HLT 2019*, pages 3795–3805. https://doi.org/10.18653/v1/N19-1380

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gökhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of NAACL-HLT 2018: Industry Papers*, pages 41–51. https://doi.org/10.18653/v1/N18-3006

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of AAAI 2020*, pages 8854–8861. https://doi.org/10.1609/aaai.v34i05.6414

Raymond Hendy Susanto and Wei Lu. 2017. Neural architectures for multilingual semantic parsing. In *Proceedings of ACL 2017*, pages 38–44.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (Almost) zero-shot cross-lingual spoken language understanding. In *Proceedings of ICASSP 2018*, pages 6034–6038.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118. https://doi.org/10.1093/llc/fqt031

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of EMNLP 2020*, pages 5052–5063.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL-HLT 2021*, pages 483–498.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of ACL 2020, System Demonstrations*, pages 87–94. https://doi.org/10.18653/v1/2020.acl-demos.12

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. In *Proceedings of the IEEE*, 101(5):1160–1179. https://doi.org/10.1109/JPROC.2012.2225812

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117. https://doi.org/10.18653/v1/2020.nlp4convai-1.13

Wei-Nan Zhang, Zhigang Chen, Wanxiang Che, Guoping Hu, and Ting Liu. 2017. The first

evaluation of Chinese human-computer dialogue technology. *arXiv preprint arXiv:1709.10217*.

Zhichang Zhang, Zhenwen Zhang, Haoyuan Chen, and Zhiman Zhang. 2019. A joint learning framework with BERT for spoken language understanding. *IEEE Access*, 7:168849–168858. https://doi.org/10.1109/ACCESS.2019.2954766

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *TACL*, 8:281–295. https://doi.org/10.1162/tacl_a_00314

Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. 2021. AllWOZ: Towards multilingual task-oriented dialog systems for all. *arXiv preprint arXiv:2112.08333*.