the SEMPRE framework (Berant et al., 2013).

ATIS has been previously translated into Chinese and Indonesian for the study of semantic parsing into λ−calculus logical forms (Susanto and Lu, 2017a), however Overnight exists only in English. To the best of our knowledge, there is presently no multi-domain dataset for executable semantic parsing in more than two languages. As previously mentioned in Section 4 , we consider Chinese and German in this paper to contrast between a language similar and dissimilar to English and also due to the reported availability of crowd-sourced workers for translation (Pavlick et al., 2014) and bilingual native speakers for verification.

To facilitate task evaluation in all languages of interest, we require a full parallel translation of ATIS in German, for comparison to the existing Chinese implementation, and a partial translation of Overnight in both German and Chinese. For task evaluation in all languages, we require a full parallel translation of ATIS to complement the existing Chinese translation from (Susanto and Lu, 2017a). As previously discussed, we translate only the development and test set of Overnight (Wang et al., 2015) into Chinese and German for assessment of crosslingual semantic parsing in a multi-domain setting. Therefore, we translate all 5,473 utterances in ATIS and 4,311 utterances in Overnight. The original Overnight dataset did not correct spelling errors from collected English paraphrases, however, we consider it unreasonable to ask participants in our task to translate misspelled words, as ambiguity in correction could lead to inaccurate translations. We subsequently identified and corrected spelling errors using word processing software.

We use Amazon Mechanical Turk (MTurk) to solicit three translations per English source sentence from crowdsourced workers (Turkers), under the assumption that this will collect at least one adequate translation (Callison-Burch, 2009). Our task design largely followed practices for translation without expert labels on MTurk (Zaidan and Callison-Burch, 2011; Post et al., 2012; Behnke et al., 2018; Sosoni et al., 2018). The task solicits translations by asking a Turker to translate 10 sentences and answer demographic questions concerning country of origin and native language. Submissions were restricted to Turkers from Germany, Austria and Switzerland or China, Singapore, and the USA for German and Chinese respectively. We built an AMT interface with quality controls which restricted Turkers from inputting whitespace and disabled copy/paste anywhere within the webpage. Attempting to copy or paste in the submission window triggered a warning that using online translation tools will result in rejection. Inauthentic translations were rejected if they held an >80% average BLEU to reference translations from Google Translate (Wu et al., 2016), as were nonsensical or irrelevant submissions. For the Chinese data collection, we also rejected submissions using Traditional Chinese Characters or Pinyin romanization. Instructions for the initial candidate collection task are given in Figure 4 and the ranking task in Figure 5. We found 94% of workers completed the optional demographic survey and that all workers reported their first language Chinese or German as desired. For Chinese, 94% of workers came from the USA and reported to have spoken Chinese for >20 years, and remaining workers resided in China. For German, all workers came from Germany and had spoken German for >25 years.

Turkers submitted 10 translations per task for $0.7 and $0.25 to rank 10 candidate translations, at an average rate to receive an equivalent full-time wage of $8.23/hour. This is markedly above the average wage for US workers of $3.01/hour discovered by Hara et al. (2019). To ensure data quality and filter disfluencies or personal biases from Turkers, we then recruited bilingual postgraduate students, native speakers of the task language, to judge if the best chosen translation from Turk was satisfactory or required rewriting. If an annotator was dissatisfied with the translation ranked best from Turk then they provided their own, which only occurred for 3.2% of all translations. Verifiers preferred the MT candidate over the Turk submissions for 29.5% of German rankings and 22.6% of Chinese rankings, however, this preference bias arose only in translations of small sentences (five or fewer words) where MT and the Turk translation were practically identical. We paid $12 an hour for this verification but to minimize cost, we did not collect multiple judgments per translation. We found that verification was completed at a rate of 60 judgments per hour, leading to an approximate cost of $2200 per language for Overnight and $2500 for ATIS into German. While this may be considered expensive, this is the minimum cost to permit comparable evaluation in every language. Sample translations for ATIS into German are given in Table 6 and sample translations for Overnight into

German and Chinese are given in Table 7.

Machine Translation In this work, we evaluate the feasibility of using machine translation (MT) as a proxy to generate in-language training data for semantic parsing of two languages. All MT systems are treated as black-box models without inspection of underlying translation mechanics or recourse for correction. For most experiments in this work, we use translations from English to the target language using Google Translate (Wu et al., 2016). We use this system owing to the purported translation quality (Duong et al., 2017) and because the API is publicly available, contrasting to the closed MT used in Conneau et al. (2018).

Additionally, we explore two approaches to modeling an ensemble of translations from multiple MT sources. We expect, but cannot guarantee, that each MT system will translate each utterance differently for greater diversity in the training corpus overall. For this approach, we consider two additional MT systems each for Chinese and German. For Mandarin, we use Baidu Translate and Youdao Translate. For German, we use Microsoft Translator Text and Yandex Translate. To verify that the ensemble of multiple MT systems provides some additional diversity, we measure the corpus level BLEU between training utterances from each source. These scores for ATIS, with comparison to human translation, and Overnight are detailed in Table 4.

Overall, we find that each MT system provides a different set of translations, with no two translation sets more similar than any other. We also find that for ATIS in German, Wu et al. (2016) provides the most similar training dataset to the gold training data. However, we find that Microsoft Translator Text appears to narrowly improve translation into Chinese by $+0.021$ BLEU. This arises as an effect of a systematic preference for a polite form of Chinese question, beginning with "请" [please], preferred by the professional translator. Overall, we collected all training data using MT for $<$ \$50 across both datasets and languages.

Translate all 10 sentences into Simplified Chinese

In this task, we ask you to provide a translation into Simplified Chinese of an English question.
You must be **native speaker of Chinese (Mandarin) and proficient in English** to complete this HIT.
We ask you to use **only Simplified Chinese characters** (简体汉字) and **do not use Pinyin** (汉语拼音).
Attempt to translate every word into Chinese. If this is difficult for rare words you do not understand, such as a person's name or place names, then please copy the English word into the translation.
You can assume all currency amounts are US Dollars and all measurements are in feet and inches.
In order to receive payment, you must complete all translations without using online translation services.
The use of online translation websites or software will be considered cheating.
Identified cheating will result in withheld payment and a ban on completing further HITs.
The demographic questionnaire is optional and you are welcome to complete as many HITs as you like.

Figure 4: Instructions provided to Turkers for the English to Chinese translation task of Overnight (Wang et al., 2015). We specify the requirement to answer in Simplified Chinese characters and specify the basis for rejection of submitted work. Instructions are condensed for brevity.

Select the best German translation for 10 English sentences

In this HIT, you will be presented with an English question and three candidate translations of this English sentence in German. We ask you to use your judgment as a native-speaker of German to select the best German translation from the three candidates.
If you consider all candidate translations to be inadequate, then provide your own translation.
You must be **native speaker of German and proficient in English** to complete this HIT.
We consider the best translation as one which asks the same question in the style of a native speaker of German, rather than the best direct translation of English. Occasionally, multiple candidates will be very similar, or identical, in this case select the first identical candidate.
You must complete all 10 to submit the HIT and receive payment.
You are welcome to submit as many HITs as you like.

Figure 5: Instructions provided to Turkers for the English to German translation ranking for both ATIS (Dahl et al., 1994) and Overnight(Wang et al., 2015). Instructions are condensed for brevity.

English	Translation into German
What ground transportation is available from the Pittsburgh airport to the town?	Welche Verkehrs Anbindung gibt es vom Pittsburgh Flughafen in die Stadt?
Could you please find me a nonstop flight from Atlanta to Baltimore on a Boeing 757 arriving at 7pm?	Könntest du für mich bitte einen Direktflug von Atlanta nach Baltimore auf einer Boeing 757 um 19 Uhr ankommend finden?
What is fare code QO mean?	Was bedeutet der ticketpreiscode QO?
Show me the cities served by Canadian Airlines International.	Zeige mir die Städte, die von den Canadian Airlines International angeflogen werden.
Is there a flight tomorrow morning from Columbus to Nashville?	Gibt es einen Flug morgen früh von Columbus nach Nashville?
Is there a Continental flight leaving from Las Vegas to New York nonstop?	Gibt es einen Continental-flug ohne Zwischenstopps, der von Las Vegas nach New York fliegt?
I would like flight information from Phoenix to Denver.	Ich hätte gerne Informationen zu Flügen von Phoenix nach Denver.
List flights from Indianapolis to Memphis with fares on Monday.	Liste Flüge von Indianapolis nach Memphis am Montag inklusive ticketpreisen auf.
How about a flight from Milwaukee to St. Louis that leaves Monday night?	Wie wäre es mit einem Flug von Milwaukee nach St. Louis, der Montag Nacht abfliegt?
A flight from St. Louis to Burbank that leaves Tuesday afternoon.	Einen Flug von St. Louis nach Burbank, der Dienstag Nachmittag abfliegt.

Table 6: Sample translations from English to German for the ATIS dataset (Dahl et al., 1994).