Meta-Learning a Cross-lingual Manifold for Semantic Parsing

Tom Sherborne and Mirella Lapata
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
tom.sherborne@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Localizing a semantic parser to support new languages requires effective cross-lingual generalization. Recent work has found success with machine-translation or zero-shot methods, although these approaches can struggle to model how native speakers ask questions. We consider how to effectively leverage minimal annotated examples in new languages for few-shot cross-lingual semantic parsing. We introduce a first-order meta-learning algorithm to train a semantic parser with maximal sample efficiency during cross-lingual transfer. Our algorithm uses high-resource languages to train the parser and simultaneously optimizes for cross-lingual generalization to lower-resource languages. Results across six languages on ATIS demonstrate that our combination of generalization steps yields accurate semantic parsers sampling $\leq$10% of source training data in each new language. Our approach also trains a competitive model on Spider using English with generalization to Chinese similarly sampling $\leq$10% of training data.[1]

1 Introduction

A semantic parser maps natural language (NL) utterances to logical forms (LF) or executable programs in some machine-readable language (e.g., SQL). Recent improvement in the capability of semantic parsers has focused on domain transfer within English (Su and Yan, 2017; Suhr et al., 2020), compositional generalization (Yin and Neubig, 2017; Herzig and Berant, 2021; Scholak et al., 2021), and, more recently, cross-lingual methods (Duong et al., 2017; Susanto and Lu, 2017b; Richardson et al., 2018).

Within cross-lingual semantic parsing, there has been an effort to bootstrap parsers with minimal data to avoid the cost and labor required to support new languages. Recent proposals include using machine translation to approximate training data for supervised learning (Moradshahi et al., 2020; Sherborne et al., 2020; Nicosia et al., 2021) and zero-shot models, which engineer cross-lingual similarity with auxiliary losses (van der Goot et al., 2021; Yang et al., 2021; Sherborne and Lapata, 2022). These shortcuts bypass costly data annotation but present limitations such as ''translationese'' artifacts from machine translation (Koppel and Ordan, 2011) or undesirable domain shift (Sherborne and Lapata, 2022). However, annotating a minimally sized data sample can potentially overcome these limitations while incurring significantly reduced costs compared to full dataset translation (Garrette and Baldridge, 2013).

We argue that a few-shot approach is more realistic for an engineer motivated to support additional languages for a database—as one can rapidly retrieve a high-quality sample of translations and combine these with existing supported languages (i.e., English). Beyond semantic parsing, cross-lingual few-shot approaches have also succeeded at leveraging a small number of annotations within a variety of tasks (Zhao et al., 2021, *inter alia*) including natural language inference, paraphrase identification, part-of-speech-tagging, and named-entity recognition. Recently, the application of meta-learning to domain generalization has further demonstrated capability for models to adapt to new domains with small samples (Gu et al., 2018; Li et al., 2018; Wang et al., 2020b).

In this work, we synthesize these directions into a meta-learning algorithm for cross-lingual semantic parsing. Our approach explicitly optimizes for cross-lingual generalization using fewer training samples per new language without performance degradation. We also require minimal

[1] Our code and data are available at github.com/tomsherborne/xgr.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 49–67, 2023. https://doi.org/10.1162/tacl_a_00533
Action Editor: Wei Lu. Submission batch: 7/2022; Revision batch: 9/2022; Published 1/2023.

computational overhead beyond standard gradient-descent training and no external dependencies beyond in-task data and a pre-trained encoder. Our algorithm, **Cross**-Lingual **G**eneralization **Reptile** (XG-REPTILE) unifies two-stage meta-learning into a single process and outperforms prior and constituent methods on all languages, given identical data constraints. The proposed algorithm is still model-agnostic and applicable to more tasks requiring sample-efficient cross-lingual transfer.

Our innovation is the combination of both intra-task and inter-language steps to jointly learn the parsing task and optimal cross-lingual transfer. Specifically, we interleave learning the overall task from a high-resource language and learning cross-lingual transfer from a minimal sample of a lower-resource language. Results on ATIS (Hemphill et al., 1990) in six languages (English, French, Portuguese, Spanish, German, Chinese) and Spider (Yu et al., 2018) in two languages (English, Chinese) demonstrate our proposal works in both single- and cross-domain environments. Our contributions are as follows:

- We introduce XG-REPTILE, a first-order meta-learning algorithm for cross-lingual generalization. XG-REPTILE approximates an *optimal manifold* using support languages with *cross-lingual regularization* using target languages to train for explicit cross-lingual similarity.
- We showcase sample-efficient cross-lingual transfer within two challenging semantic parsing datasets across multiple languages. Our approach yields more accurate parsing in a few-shot scenario and demands $10\times$ fewer samples than prior methods.
- We establish a cross-domain and cross-lingual parser obtaining promising results for both Spider in English (Yu et al., 2018) and CSpider in Chinese (Min et al., 2019).

2 Related Work

Meta-Learning for Generalization Meta-Learning[2] has recently emerged as a promising technique for generalization, delivering high performance on unseen domains by *learning to learn*, that is, improving learning over multiple episodes (Hospedales et al., 2022; Wang et al., 2021b). A popular approach is Model-Agnostic Meta-Learning (Finn et al., 2017, MAML), wherein the goal is to train a model on a variety of learning tasks, such that it can solve *new* tasks using a small number of training samples. In effect, MAML facilitates task-specific fine-tuning using few samples in a two-stage process. MAML requires computing higher-order gradients (i.e., ''gradient through a gradient'') which can often be prohibitively expensive for complex models. This limitation has motivated *first-order* approaches to MAML which offer similar performance with improved computational efficiency.

In this vein, the Reptile algorithm (Nichol et al., 2018) transforms the higher-order gradient approach into K successive first-order steps. Reptile-based training approximates a solution *manifold* across tasks (i.e., a high-density parameter sub-region biased for strong cross-task likelihood), then similarly followed by rapid fine-tuning. By learning an optimal initialization, meta-learning proves useful for low-resource adaptation by minimizing the data required for out-of-domain tuning on new tasks. Kedia et al. (2021) also demonstrate the utility of Reptile to improve *single-task* performance. We build on this to examine single-task cross-lingual transfer using the *manifold* learned with Reptile.

Meta-Learning for Semantic Parsing A variety of NLP applications have adopted meta-learning in zero- and few-shot learning scenarios as a method of explicitly training for generalization (Lee et al., 2021; Hedderich et al., 2021). Within semantic parsing, there has been increasing interest in *cross-database generalization*, motivated by datasets such as Spider (Yu et al., 2018) requiring navigation of unseen databases (Herzig and Berant, 2017; Suhr et al., 2020).

Approaches to generalization have included simulating source and target domains (Givoli and Reichart, 2019) and synthesizing new training data based on unseen databases (Zhong et al., 2020; Xu et al., 2020a). Meta-learning has demonstrated fast adaptation to new data within a monolingual low-resource setting (Huang et al., 2018; Guo et al., 2019; Lee et al., 2019; Sun et al., 2020). Similarly, Chen et al. (2020) utilize Reptile to improve generalization of a model, trained on source domains, to fine-tune on new domains.

[2]We refer the interested reader to Wang et al. (2020b), Hospedales et al. (2022), and Wang et al. (2021b) for more extensive surveys on meta-learning.

Our work builds on Wang et al. (2021a), who explicitly promote monolingual cross-domain generalization by ''meta-generalizing'' across disjoint, domain-specific batches during training.

Cross-lingual Semantic Parsing A surge of interest in cross-lingual NLU has seen the creation of many benchmarks across a breadth of languages (Conneau et al., 2018; Hu et al., 2020; Liang et al., 2020), thereby motivating significant exploration of cross-lingual transfer (Nooralahzadeh et al., 2020; Xia et al., 2021; Xu et al., 2021; Zhao et al., 2021, *inter alia*). Previous approaches to cross-lingual semantic parsing assume parallel multilingual training data (Jie and Lu, 2014) and exploit multi-language inputs for training without resource constraints (Susanto and Lu, 2017a,b).

There has been recent interest in evaluating if machine translation is an economic proxy for creating training data in new languages (Sherborne et al., 2020; Moradshahi et al., 2020). Zero-shot approaches to cross-lingual parsing have also been explored using auxiliary training objectives (Yang et al., 2021; Sherborne and Lapata, 2022). Cross-lingual learning has also been gaining traction in the adjacent field of spoken-language understanding (SLU). For datasets such as Multi-ATIS (Upadhyay et al., 2018), MultiATIS++ (Xu et al., 2020b), and MTOP (Li et al., 2021), zero-shot cross-lingual transfer has been studied through specialized decoding methods (Zhu et al., 2020), machine translation (Nicosia et al., 2021), and auxiliary objectives (van der Goot et al., 2021).

Cross-lingual semantic parsing has mostly remained orthogonal to the cross-database generalization challenges raised by datasets such as Spider (Yu et al., 2018). While we primarily present findings for multilingual ATIS into SQL (Hemphill et al., 1990), we also train a parser on both Spider and its Chinese version (Min et al., 2019). To the best of our knowledge, we are the first to explore a multilingual approach to this cross-database benchmark. We use Reptile to learn the overall task and leverage domain generalization techniques (Li et al., 2018; Wang et al., 2021a) for sample-efficient cross-lingual transfer.

3 Problem Definition

Semantic Parsing We wish to learn a parameterized parsing function, p_θ, which maps from a natural language utterance and a relational database context to an executable program expressed in a logical form (LF) language:

$$P = p_\theta\left(Q,\ D\right) \tag{1}$$

As formalized in Equation (1), we learn parameters, θ, using paired data $(Q, P, \mathcal{D})$ where P is the logical form equivalent of natural language question Q. In this work, our LFs are all executable SQL queries and therefore grounded in a database $\mathcal{D}$. A single-domain dataset references only one $\mathcal{D}$ database for all (Q, P), whereas a multi-domain dataset demands reasoning about unseen databases to generalize to new queries. This is expressed as a 'zero-shot' problem if the databases at test time, $\mathcal{D}_{\text{test}}$, were unseen during training. This challenge demands a parser capable of *domain generalization* beyond observed databases. This is in addition to the *structured prediction* challenge of semantic parsing.

Cross-Lingual Generalization Prototypical semantic parsing datasets express the question, Q, in English only. As discussed in Section 1, our parser should be capable of mapping from *additional* languages to well-formed, executable programs. However, prohibitive expense limits us from reproducing a monolingual model for each additional language and previous work demonstrates accuracy improvement by training *multilingual* models (Jie and Lu, 2014). In addition to the challenges of structured prediction and domain generalization, we jointly consider *cross-lingual generalization*. Training primarily relies on existing English data (i.e., Q_{EN} samples) and we show that our meta-learning algorithm in Section 4 leverages a small sample of training data in new languages for accurate parsing. We express this sample, $\mathcal{S}_l$, for some language, l, as:

$$\mathcal{S}_l = (Q_l,\ P,\ \mathcal{D})_{i=0}^{N_l} \tag{2}$$

where N_l is the sample size from l, assumed to be smaller than the original English dataset (i.e., $N_l \ll N_{\text{EN}}$). Where available, we extend this paradigm to develop models for L different languages simultaneously in a multilingual setup by combining samples as:

$$\mathcal{S}_{\text{L}} = \{\mathcal{S}_{l_1}, \mathcal{S}_{l_2}, \ldots, \mathcal{S}_{l_N}\} \tag{3}$$

We can express cross-lingual generalization as:

$$p_\theta\left(P \mid Q_l,\ \mathcal{D}\right) \rightarrow p_\theta\left(P \mid Q_{\text{EN}},\ \mathcal{D}\right) \quad (4)$$

where $p_\theta\left(P \mid Q_{\text{EN}},\ \mathcal{D}\right)$ is the predicted distribution over all possible output SQL sequences conditioned on an English question, Q_{EN}, and a database $\mathcal{D}$. Our goal is for the prediction from a new language, Q_l, to converge towards this existing distribution using the same parameters θ, constrained to fewer samples in l than English.

We aim to maximize the accuracy of predicting programs on unseen test data from each non-English language l. The key challenge is learning a performant distribution over each new language with minimal available samples. This includes learning to incorporate each l into the parsing task and modeling the language-specific surface form of questions. Our setup is akin to few-shot learning; however, the number of examples needed for satisfactory performance is an empirical question. We are searching for both minimal sample sizes and maximal sampling efficiency. We discuss our sampling strategy in Section 5.2 with results at multiple sizes of $\mathcal{S}_L$ in Section 6.

4 Methodology

We combine two meta-learning techniques for cross-lingual semantic parsing. The first is the Reptile algorithm outlined in Section 2. Reptile optimizes for dense likelihood regions within the parameters (i.e., a solution *manifold*) through promoting inter-batch generalization (Nichol et al., 2018). Standard Reptile iteratively optimizes the manifold for an improved initialization across objectives. Rapid fine-tuning yields the final task-specific model. The second technique is the first-order approximation of DG-MAML (Li et al., 2018; Wang et al., 2021a). This single-stage process optimizes for domain generalization by simulating ‘‘source’’ and ‘‘target’’ batches from different domains to explicitly optimize for *cross-batch* generalization. Our algorithm, XG-Reptile, combines these paradigms to optimize a target loss with the overall learning ‘‘direction’’ derived as the *optimal manifold* learned via Reptile. This trains an accurate parser demonstrating sample-efficient cross-lingual transfer within an efficient *single-stage* learning process.

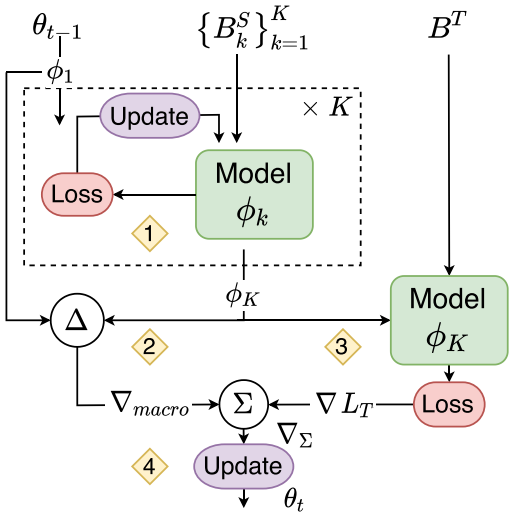


Figure 1: One iteration of XG-Reptile. (1) Run K iterations of gradient descent over K support batches to learn ϕ_K, (2) compute ∇_{macro}, the difference between ϕ_K and ϕ_1, (3) find the loss on the target batch using ϕ_K, and (4) compute the final gradient update from ∇_{macro} and the target loss.

4.1 The XG-Reptile Algorithm

Each learning episode of XG-Reptile comprises two component steps: *intra-task* learning and *inter-language* generalization to jointly learn parsing and cross-lingual transfer. Alternating these processes trains a competitive parser from multiple languages with low computational overhead beyond existing gradient-descent training. Our approach combines the typical two stages of meta-learning to produce a single model without a fine-tuning requirement.

Task Learning Step We first sample from the high-resource language (i.e., $\mathcal{S}_{\text{EN}}$) K ‘‘support’’ batches of examples, $\mathcal{B}^S = \{(Q_{\text{EN}}, P, \mathcal{D})\}$. For each of K batches: We compute predictions, compute losses, calculate gradients and adjust parameters using some optimizer (see illustration in Figure 1). After K successive optimization steps the initial weights in this episode, ϕ_1, have been optimized to ϕ_K. The difference between final and initial weights is calculated as:

$$\nabla_{macro} = \phi_K - \phi_1 \quad (5)$$

This ‘‘macro-gradient’’ step is equivalent to a Reptile step (Nichol et al., 2018), representing learning a *solution manifold* as an approximation of overall learning trajectory.

Algorithm 1 XG-REPTILE

Require: Support data, $\mathcal{S}_{\text{EN}}$, target data, $\mathcal{S}_L$
Require: Inner learning rate, α, outer learning rate, β
1: Initialise θ_1, the vector of initial parameters
2: **for** $t \leftarrow 1$ **to** T **do**
3: Copy $\phi_1 \leftarrow \theta_{t-1}$
4: Sample K support batches $\{\mathcal{B}^S\}_{k=1}^{K}$ from $\mathcal{S}_{\text{EN}}$
5: Sample target language l from L languages
6: Sample target batch $\mathcal{B}^T$ from $\mathcal{S}_l$
7: **for** $k \leftarrow 1$ **to** K [Inner Loop] **do**
8: $\mathcal{L}_k^S \leftarrow \text{Forward}\left(\mathcal{B}_k^S,\ \phi_{k-1}\right)$
9: $\phi_k \leftarrow \text{Adam}\left(\phi_{k-1},\ \nabla\mathcal{L}_k^S,\ \alpha\right)$
10: **end for**
11: Macro grad: $\nabla_{macro} \leftarrow \phi_K - \phi_1$
12: Target Step: $\mathcal{L}_T \leftarrow \text{Forward}\left(\mathcal{B}^T,\ \phi_K\right)$
13: Total gradient: $\nabla_\Sigma = \nabla_{macro} + \nabla_{\phi_K}\mathcal{L}_T$
14: Update $\theta_t \leftarrow \text{SGD}\left(\theta_{t-1},\ \nabla_\Sigma,\ \beta\right)$
15: **end for**

Cross-Lingual Step The second step samples one ''target'' batch, $\mathcal{B}^T = (Q_l,\ P,\ \mathcal{D})$, from a sampled target language (i.e., $\mathcal{S}_l \subset \mathcal{S}_{\text{L}}$). We compute the cross-entropy loss and gradients from the prediction of the model at ϕ_K on $\mathcal{B}^T$:

$$\mathcal{L}_T = \text{Loss}\left(p_{\phi_K}\left(Q_l, \mathcal{D}\right), P\right) \tag{6}$$

We evaluate the parser at ϕ_K on a target language we desire to generalize to. We show below that the gradient of $\mathcal{L}_T$ comprises the loss at ϕ_K and additional terms maximizing the inner product between the high-likelihood manifold and the target loss. The total gradient encourages intra-task and cross-lingual learning (see Figure 1).

Algorithm 1 outlines the XG-REPTILE process (loss calculation and batch processing are simplified for brevity). We repeat this process over T episodes to train model p_θ to convergence. If we optimized for target data to align with individual support batches (i.e., $K = 1$) then we may observe batch-level noise in cross-lingual generalization. Our intuition is that aligning the target gradient with an approximation of the task manifold, i.e., ∇_{macro}, will overcome this noise and align new languages to a more mutually beneficial direction during training. We observe this intuitive behavior during learning in Section 6.

We efficiently generalize to low-resource languages by exploiting the asymmetric data requirements between steps: One batch of the target language is required for K batches of the source language. For example, if $K = 10$ then using this $\frac{1}{K}$ proportionality requires 10% of target-language data relative to support. We demonstrate in Section 6 that we can use a smaller $< \frac{1}{K}$ quantity per target language to increase sample efficiency.

Gradient Analysis Following Nichol et al. (2018), we express $g_k = \nabla\mathcal{L}_k^S$, the gradient in a single step of the inner loop (Line 9), as:

$$g_k = \bar{g}_i + \bar{H}_k\left(\phi_k - \phi_1\right) + O\left(\alpha^2\right) \tag{7}$$

We use a Taylor series expansion to approximate g_k by $\bar{g}_k$, the gradient at the original point, ϕ_1, the Hessian matrix of the gradient at the initial point, $\bar{H}_k$, the step difference between position ϕ_k and the initial position and some scalar terms with marginal influence, $O\left(\alpha^2\right)$.

By evaluating Equation (7) at $i = 1$ and rewriting the difference as a sum of gradient steps (e.g., Equations (8) and (9)), we arrive at an expression for g_k shown in Equation (10) expressing the gradient as an initial component, $\hat{g}_k$, and the product of the Hessian at k, with all prior gradient steps. We refer to Nichol et al. (2018) for further validation that the gradient of this product maximizes the cross-batch expectation—therefore promoting cross-batch generalization and towards the solution manifold. The final gradient (Equation (11)) is the accumulation over g_k steps and is equivalent to Equation (5). ∇_{macro} comprises both gradients of K steps and additional terms maximizing the inner-product of cross-batch gradients.

$$\text{Use}\ \ g_j = \bar{g}_j + O(\alpha) \tag{8}$$

$$\phi_k - \phi_1 = -\alpha\sum_{j=1}^{k-1} g_j \tag{9}$$

$$g_k = \bar{g}_i - \alpha\bar{H}_i\sum_{j=1}^{k-1}\bar{g}_j + O\left(\alpha^2\right) \tag{10}$$

$$\nabla_{macro} = \sum_{k=1}^{K} g_k \tag{11}$$

We can similarly express the gradient of the target batch as Equation (12) where the term, $\bar{H}_T\nabla_{macro}$, is the cross-lingual generalization product similar to the intra-task generalization seen above.

$$g_T = \bar{g}_T - \alpha\bar{H}_T\nabla_{macro} + O\left(\alpha^2\right) \tag{12}$$

Equation (13) shows an example final gradient when $K = 2$. Within the parentheses are the cross-batch and cross-lingual gradient products as components promoting fast learning across multiple axes of generalization.

$$\begin{aligned}\nabla_{\Sigma} &= g_1 + g_2 + g_T \\ &= \bar{g}_1 + \bar{g}_2 + \bar{g}_T \\ &- \alpha\left(\bar{H}_2\bar{g}_1 + \bar{H}_T[\bar{g}_1 + \bar{g}_2]\right) + O\left(\alpha^2\right)\end{aligned} \tag{13}$$

The key hyperparameter in XG-REPTILE is the number of inner-loop steps K representing a trade-off between manifold approximation and target step frequency. At small K, the manifold approximation may be poor, leading to suboptimal learning. At large K, then improved manifold approximation incurs fewer target batch steps per epoch, leading to weaked cross-lingual transfer. In practice, K is set empirically, and Section 6 identifies an optimal region for our task.

XG-REPTILE can be viewed as generalizing two existing algorithms. Without the $\mathcal{L}_T$ loss, our approach is equivalent to Reptile and lacks cross-lingual alignment. If $K = 1$, then XG-REPTILE is equivalent to DG-FMAML (Wang et al., 2021a) but lacks generalization across support batches. Our unification of these algorithms represent the best of both approaches and outperforms both techniques within semantic parsing. Another perspective is that XG-REPTILE learns a *regularized manifold*, with immediate cross-lingual capability, as opposed to standard Reptile, which requires fine-tuning to transfer across tasks. We identify how this contrast in approaches influences cross-lingual transfer in Section 6.

5 Experimental Design

We evaluate XG-REPTILE against several comparison systems across multiple languages. Where possible, we re-implement existing models and use identical data splits to isolate the contribution of our training algorithm.

5.1 Data

We report results on two semantic parsing datasets. First on **ATIS** (Hemphill et al., 1990), using the multilingual version from Sherborne and Lapata (2022) pairing utterances in six languages (English, French, Portuguese, Spanish, German, Chinese) to SQL queries. ATIS is split into 4,473 training pairs with 493 and 448 examples for validation and testing, respectively. We report performance as execution accuracy to test if predicted SQL queries can retrieve accurate database results.

We also evaluate on **Spider** (Yu et al., 2018), combining English and Chinese (Min et al., 2019, CSpider) versions as a cross-lingual task. The latter translates all questions to Chinese but retains the English database. Spider is significantly more challenging; it contains 10,181 questions and 5,693 unique SQL queries for 200 multi-table databases over 138 domains. We use the same split as Wang et al. (2021a) to measure generalization to unseen databases/table-schema during testing. This split uses 8,659 examples from 146 databases for training and 1,034 examples from 20 databases for validation. The test set contains 2,147 examples from 40 held-out databases and is held privately by the authors. To our knowledge, we report the first multilingual approach for Spider by training one model for English and Chinese. Our challenge is now multi-dimensional, requiring cross-lingual and cross-domain generalization. Following Yu et al. (2018), we report exact set match accuracy for evaluation.

5.2 Sampling for Generalization

Training for cross-lingual generalization often uses parallel samples across languages. We illustrate this in Equation (14), where y_1 is the equivalent output for inputs, x_1, in each language:

$$\text{EN} : (x_1, y_1)\ \text{DE} : (x_1, y_1)\ \text{ZH} : (x_1, y_1) \tag{14}$$

However, high sample overlap risks trivializing the task because models are not learning from new pairs, but instead matching only new *inputs* to known outputs. A preferable evaluation will test composition of novel outputs from unseen inputs:

$$\text{EN} : (x_1, y_1)\ \ \text{DE} : (x_2, y_2)\ \ \text{ZH} : (x_2, y_2) \tag{15}$$

Equation (15) samples exclusive, disjoint datasets for English and target languages during training. In other words, this process is *subtractive*—for example, a 5% sample of German (or Chinese) target data leaves 95% of data as the English support. This is similar to K-fold cross-validation used to evaluate across many data splits. We

sample data for our experiments with Equation (15). It is also possible to use Equation (16), where target samples are also disjoint, but we find this setup results in too few English examples for effective learning.

$$\text{EN} : (x_1, y_1)\ \text{DE} : (x_2, y_2)\ \text{ZH} : (x_3, y_3) \quad (16)$$

5.3 Semantic Parsing Models

We use a Transformer encoder-decoder model similar to Sherborne and Lapata (2022) for our ATIS experiments. We use the same mBART50 encoder (Tang et al., 2021) and train a Transformer decoder from scratch to generate SQL.

For Spider, we use the RAT-SQL model (Wang et al., 2020a), which has formed the basis of many performant submissions to the Spider leaderboard. RAT-SQL can successfully reason about unseen databases and table schema using a novel schema-linking approach within the encoder. We use the version from Wang et al. (2021a) with mBERT (Devlin et al., 2019) input embeddings for a unified model between English and Chinese inputs. Notably, RAT-SQL can be over-reliant on lexical similarity features between input questions and tables (Wang et al., 2020a). This raises the challenge of generalizing to Chinese where such overlap is null. For fair comparison, we implement identical models as prior work on each dataset and only evaluate the change in training algorithm. This is why we use an mBART50 encoder component for ATIS experiments and different mBERT input embeddings for Spider experiments.

5.4 Comparison Systems

We compare our algorithm against several strong baselines and adjacent training methods including:

Monolingual Training A monolingual Transformer is trained on gold-standard professionally translated data for each new language. This is a monolingual upper bound without few-shot constraints.

Multilingual Training A multilingual Transformer is trained on the union of all data from the ‘‘Monolingual Training’’ method. This ideal upper bound uses all data in all languages without few-shot constraints.

Translate-Test A monolingual Transformer is trained on source English data ($\mathcal{S}_{\text{EN}}$). Machine translation is used to translate test data from additional languages into English. Logical forms are predicted from translated data using the English model.

Translate-Train Machine translation is used to translate English training data into each target language. A monolingual Transformer is trained on translated training data and logical forms are predicted using this model.

Train-EN∪All A Transformer is trained on English data and samples from *all* target languages together in a single stage (i.e., $\mathcal{S}_{\text{EN}} \cup \mathcal{S}_L$). This is superior to training without English (e.g., on $\mathcal{S}_L$ only); we contrast to this approach for more competitive comparison.

TrainEN→FT-All We first train on English support data, $\mathcal{S}_{\text{EN}}$, and then fine-tune on target samples, $\mathcal{S}_L$.

Reptile-EN→FT-All Initial training uses Reptile (Nichol et al., 2018) on English support data, $\mathcal{S}_{\text{EN}}$, followed by fine-tuning on target samples, $\mathcal{S}_L$. This is a typical usage of Reptile for training a low-resource multi-domain parser (Chen et al., 2020).

We also compare to DG-FMAML (Wang et al., 2021a) as a special case of XG-Reptile when $K = 1$. Additionally, we omit pairwise versions of XG-Reptile (e.g., separate models generalizing from English to individual languages). These approaches demand more computation and demonstrated no significant improvement over a multi-language approach. All Machine Translation uses Google Translate (Wu et al., 2016).

5.5 Training Configuration

Experiments focus on the expansion from English to additional languages, where we use English as the ‘‘support’’ language and additional languages as ‘‘target’’. Key hyperparameters are outlined in Table 1. We train each model using the given optimizers with early stopping where model selection is through minimal validation loss for combined support and target languages. Input utterances are tokenized using SentencePiece (Kudo and Richardson, 2018) and Stanza (Qi et al., 2020) for ATIS and Spider, respectively. All experiments are implemented in PyTorch on a single

	ATIS	Spider
Batch Size	10	16
Inner Optimizer	SGD	
Inner LR	1×10^{-4}	
Outer Optimizer	Adam (Kingma and Ba, 2015)	
Outer LR	1×10^{-3}	5×10^{-4}
Optimum K	10	3
Max Train Steps	20,000	
Training Time	12 hours	2.5 days

Table 1: Experimental hyperparameters for XG-REPTILE on ATIS and Spider set primarily by replicating prior work.

V100 GPU. We report key results for ATIS averaged over three seeds and five random data splits. For Spider, we submit the best singular model from five random splits to the leaderboard.

6 Results and Analysis

We contrast XG-REPTILE to baselines for ATIS in Table 2 and present further analysis within Figure 2. Results for the multi-domain Spider are shown in Table 3. Our findings support our hypothesis that XG-REPTILE is a superior algorithm for jointly training a semantic parser and encouraging cross-lingual generalization with improved sample efficiency. Given the same data, XG-REPTILE produces more mutually beneficial parameters for both model requirements with only modifications to the training loop.

Comparison across Generalization Strategies We compare XG-REPTILE to established learning algorithms in Table 2. Across baselines, we find that single-stage training, that is, *Train-EN∪All* or machine-translation based models, perform below two-stage approaches. The strongest competitor is the *Reptile-EN→FT-All* model, highlighting the effectiveness of Reptile for single-task generalization (Kedia et al., 2021). However, XG-REPTILE performs above all baselines across sample rates. Practically, 1%, 5%, 10% correspond to 45, 225, and 450 example pairs, respectively. We identify significant improvements ($p < 0.01$; relative to the closest model using an independent t-test) in cross-lingual transfer through jointly learning to parse and multi-language generalization while maintaining single-stage training efficiency.

Compared to the upper bounds, XG-REPTILE performs above *Monolingual Training* at $\geq 1\%$ sampling, which further supports the prior benefit of multilingual modeling (Susanto and Lu, 2017a). *Multilingual Training* is only marginally stronger than XG-REPTILE at 1% and 5% sampling despite requiring many more examples. XG-REPTILE@10% improves on this model by an average +1.3%. Considering that our upper bound uses 10× the data of XG-REPTILE@10%, this accuracy gain highlights the benefit of explicit cross-lingual generalization. This is consistent at higher sample sizes (see Figure 2(c) for German).

At the smallest sample size, XG-REPTILE@1%, demonstrates a +12.4% and +13.2% improvement relative to *Translate-Train* and *Translate-Test*. Machine translation is often viable for cross-lingual transfer (Conneau et al., 2018). However, we find that mistranslation of named entities incurs an exaggerated parsing penalty—leading to inaccurate logical forms (Sherborne et al., 2020). This suggests that sample quality has an exaggerated influence on semantic parsing performance. When training XG-REPTILE with MT data, we also observe a lower Target-language average of 66.9%. This contrast further supports the importance of sample quality in our context.

XG-REPTILE improves cross-lingual generalization across all languages at equivalent and lower sample sizes. At 1%, it improves by an average +15.7% over the closest model, *Reptile-EN→FT-All*. Similarly, at 5%, we find +9.8% gain, and at 10%, we find +8.9% relative to the closest competitor. Contrasting across sample sizes—our best approach is @10%, however, this is +3.5% above @1%, suggesting that smaller samples could be sufficient if 10% sampling is unattainable. This relative stability is an improvement compared to the 17.7%, 11.2%, or 10.3% difference between @1% and @10% for other models. This implies that XG-REPTILE better utilizes smaller samples than adjacent methods.

Across languages at 1%, XG-REPTILE improves primarily for languages dissimilar to English (Ahmad et al., 2019) to better minimize the cross-lingual transfer gap. For Chinese (ZH), we see that XG-REPTILE@1% is +26.4% above the closest baseline. This contrasts with the smallest gain, +8.5% for German, with greater similarity to English. Our improvement also yields less variability across target languages—the standard

		EN	FR	PT	ES	DE	ZH	Target Avg
ZX-PARSE (Sherborne and Lapata, 2022)		76.9	70.2	63.4	59.7	69.3	60.2	64.6 ± 5.0
Monolingual Training		77.2	67.8	66.1	64.1	66.6	64.9	65.9 ± 1.4
Multilingual Training		73.9	72.5	73.1	70.4	72.0	70.5	71.7 ± 1.2
Translate-Train		—	55.9	56.1	57.1	60.1	56.1	57.1 ± 1.8
Translate-Test		—	58.2	57.3	57.9	56.9	51.4	56.3 ± 2.8
@1%	Train-EN∪All	69.7 ± 1.4	44.0 ± 3.5	42.2 ± 3.7	38.3 ± 6.8	45.8 ± 2.6	41.7 ± 3.6	42.4 ± 2.8
	Train-EN→FT-All	71.2 ± 2.3	53.3 ± 5.2	49.7 ± 5.4	56.1 ± 2.7	52.5 ± 6.7	39.0 ± 4.0	50.1 ± 6.6
	Reptile-EN→FT-All	73.2 ± 0.7	58.9 ± 4.8	54.8 ± 3.4	52.8 ± 4.4	60.6 ± 3.6	41.7 ± 4.0	53.8 ± 7.4
	XG-REPTILE	**73.8** ± 0.3	**70.4** ± 1.8	**70.8** ± 0.7	**68.9** ± 2.3	**69.1** ± 1.2	**68.1** ± 1.2	**69.5** ± 1.1
@5%	Train-EN∪All	67.3 ± 1.6	55.2 ± 4.5	54.7 ± 4.5	44.4 ± 4.5	55.8 ± 2.9	52.3 ± 4.3	52.5 ± 4.7
	Train-EN→FT-All	69.2 ± 1.9	58.9 ± 5.3	54.8 ± 5.4	52.8 ± 4.5	60.6 ± 6.5	41.7 ± 9.5	53.8 ± 7.4
	Reptile-EN→FT-All	69.5 ± 1.8	65.3 ± 3.8	61.3 ± 6.0	59.6 ± 2.6	64.9 ± 5.1	56.9 ± 9.2	61.6 ± 3.6
	XG-REPTILE	**74.4** ± 1.3	**73.0** ± 0.9	**71.6** ± 1.1	**71.6** ± 0.7	**71.1** ± 0.6	**69.5** ± 0.5	**71.4** ± 1.3
@10%	Train-EN∪All	65.7 ± 1.9	61.5 ± 1.7	62.1 ± 2.3	53.7 ± 3.2	62.7 ± 2.3	60.6 ± 2.4	60.1 ± 3.7
	Train-EN→FT-All	67.4 ± 1.9	63.8 ± 5.8	60.3 ± 5.3	59.6 ± 4.0	64.5 ± 6.5	58.4 ± 6.4	61.3 ± 2.7
	Reptile-EN→FT-All	72.8 ± 1.8	66.3 ± 4.2	64.6 ± 4.9	62.3 ± 6.4	66.6 ± 5.0	60.7 ± 3.6	64.1 ± 2.6
	XG-REPTILE	**75.8** ± 1.3	**74.2** ± 0.2	**72.8** ± 0.6	**72.1** ± 0.7	**73.0** ± 0.6	**72.8** ± 0.5	**73.0** ± 0.8

Table 2: Denotation accuracy using varying learning algorithms including XG-REPTILE at 1%, 5%, and 10% sampling rates for target dataset size relative to support dataset for ATIS. We report for *English*, *French*, *Portuguese*, *Spanish*, *German*, and *Chinese*. *Target Avg* reports the average denotation accuracy across non-English languages ± standard deviation across languages. For few-shot experiments, we also report the standard deviation (±) across random samples. Best few-shot results per language are bolded.

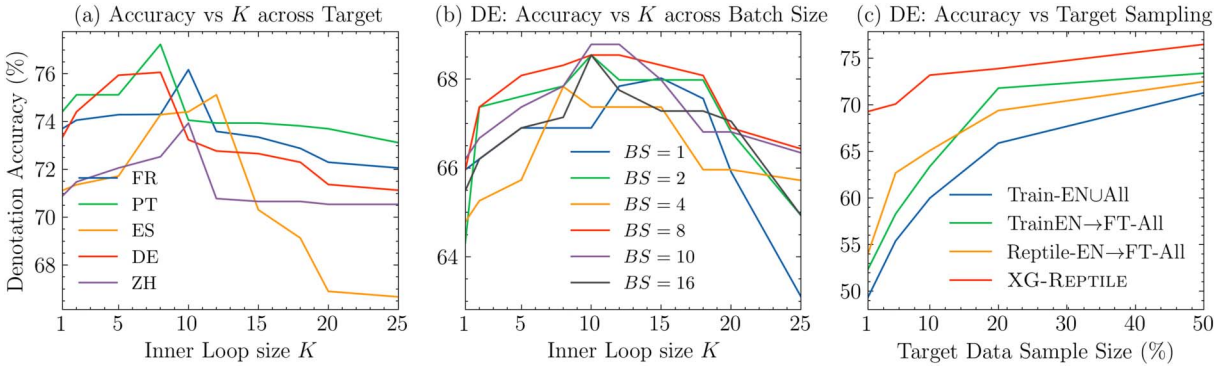


Figure 2: Ablation Experiments on ATIS (a) accuracy against inner loop size K across languages, (b) accuracy against K for German when varying batch size, and (c) accuracy against dataset sample size relative to support dataset from 1% to 50% for German. For (b), the $K = 1$ case is equivalent to DG-FMAML (Wang et al., 2021a).

deviation across languages for XG-REPTILE@1% is 1.1, compared to 2.8 for *Train-EN∪All* or 7.4 for *Reptile-EN→FT-All*.

We can also compare to ZX-PARSE, the method of Sherborne and Lapata (2022) that engineers cross-lingual latent alignment for zero-shot semantic parsing without data in target languages. With 45 samples per target language, XG-REPTILE@1% improves by an average of +4.9%. XG-REPTILE is more beneficial for distant languages—cross-lingual transfer penalty between English and Chinese is −12.3% for ZX-PARSE compared to −5.7% in our case. While these systems are not truly comparable, given different data requirements, this contrast is practically useful for comparison between zero- and few-shot localization.

Influence of K on Performance In Figure 2(a) we study how variation in the key hyperparameter K, the size of the inner-loop in Algorithm 1 or the number of batches used to approximate the *solution manifold* influences model performance across languages (single run at 5% sampling). When $K = 1$, the model learns generalization from batch-wise similarity, which is equivalent to DG-FMAML (Wang et al., 2021a). We empirically find that increasing K beyond

one benefits performance by encouraging cross-lingual generalization with the *task* over a single *batch*, and it is, therefore, beneficial to align an out-of-domain example with the overall *direction* of training. However, as theorized in Section 4, increasing K also decreases the frequency of the outer step within an epoch leading to poor cross-lingual transfer at high K. This trade-off yields an optimal operating regime for this hyper-parameter. We use $K = 10$ in our experiments as the center of this region. Given this setting of K, the target sample size must be 10% of the support sample size for training in a single epoch. However, Table 2 identifies XG-REPTILE as the most capable algorithm for ''over-sampling'' smaller target samples for resource-constrained generalization.

Influence of Batch Size on Performance We consider two further case studies to analyze XG-REPTILE performance. For clarity, we focus on German; however, these trends are consistent across all target languages. Figure 2(b) examines if the effects of cross-lingual transfer within XG-REPTILE are sensitive to batch size during training (single run at 5% sampling). A dependence between K and batch size could imply that the desired inter-task and cross-lingual generalization outlined in Equation (13) is an unrealistic, edge-case phenomenon. This is not the case, and a trend of optimal K setting is consistent across many batch sizes. This suggests that K is an independent hyper-parameter requiring tuning alongside existing experimental settings.

Performance across Larger Sample Sizes We consider a wider range of target data sample sizes between 1% and 50% in Figure 2(c). We observe that baseline approaches converge to between 69.3% and 73.9% at 50% target sample size. Surprisingly, the improvement of XG-REPTILE is retained at higher sample sizes with an accuracy of 76.5%. The benefit of XG-REPTILE is still greatest at low sample sizes with +5.4% improvement at 1%; however, we maintain a +2.6% gain over the closest system at 50%. While low sampling is the most economical, the consistent benefit of XG-REPTILE suggests a promising strategy for other cross-lingual tasks.

Learning Spider and CSpider Our results on Spider and CSpider are shown in Table 3. We compare XG-REPTILE to monolingual approaches from Wang et al. (2021a) and discuss cross-lingual results when sampling between 1% and 10% of CSpider target during training.

		EN		ZH	
		Dev	Test	Dev	Test
Monolingual					
DG-MAML		**68.9**	**65.2**	**50.4**	**46.9**
DG-FMAML		56.8	—	32.5	—
XG-REPTILE		63.5	—	48.9	—
Multilingual					
XG-REPTILE	@1%	56.8	56.5	47.0	45.6
	@5%	**59.6**	58.1	47.3	45.6
	@10%	59.2	**59.7**	**48.0**	**46.0**

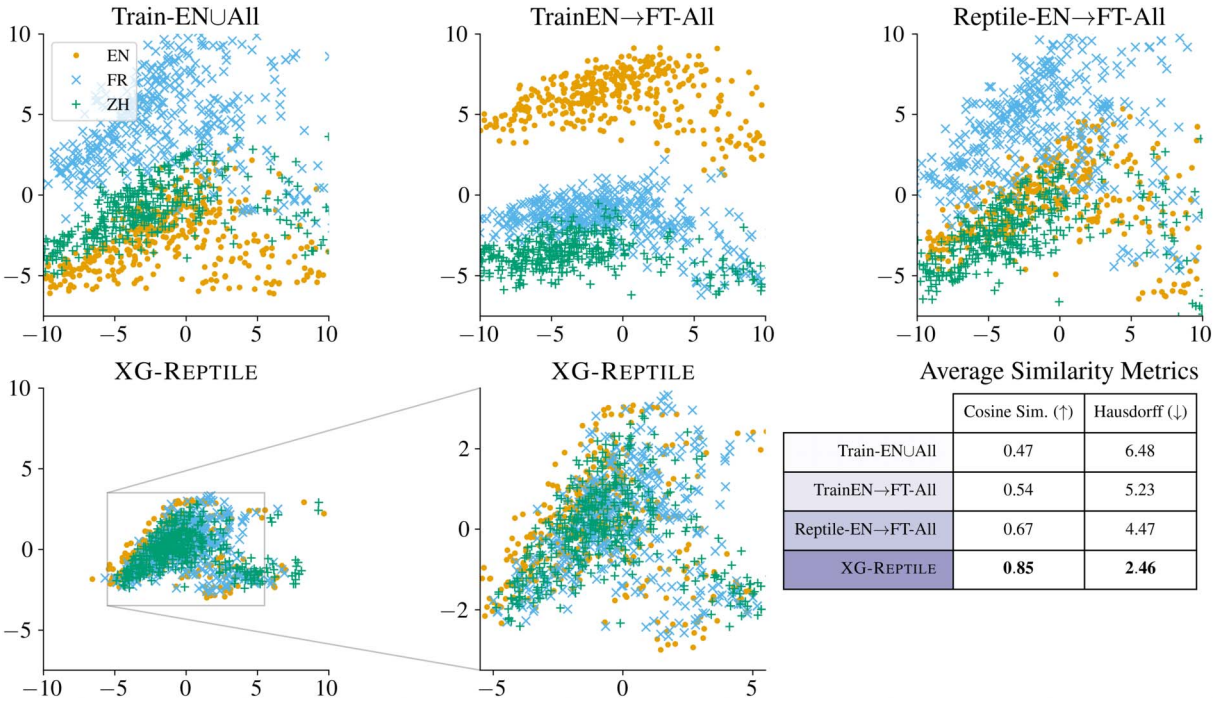
Table 3: Exact set match accuracy for RAT-SQL trained on Spider (English) and CSpider (Chinese) comparing XG-REPTILE to DG-MAML and DG-FMAML (Wang et al., 2021a). We experiment with sampling between 1% to 10% of Chinese examples relative to English. Monolingual and multilingual best results are bolded.

In the *monolingual setting*, XG-REPTILE shows significant improvement ($p < 0.01$; using an independent samples t-test) compared to DG-FMAML with +6.7% for English and +16.4% for Chinese dev accuracy. This further supports our claim that generalizing with a *task manifold* is superior to batch-level generalization.

Our results are closer to DG-MAML (Wang et al., 2021a), a higher-order meta-learning method requiring computational resources and training times exceeding $4\times$ the requirements for XG-REPTILE. XG-REPTILE yields accuracies -5.4% and -1.5% below DG-MAML for English and Chinese, where DG-FMAML performs much lower at -12.1% (EN) and -17.9% (ZH). Our results suggest that XG-REPTILE is a superior first-order meta-learning algorithm rivaling prior work with greater computational demands.[3]

In the multilingual setting, we observe that XG-REPTILE performs competitively using as little as 1% of Chinese examples. While training sampling 1% and 5% perform similarly—the best model sees 10% of CSpider samples during training to yield accuracy only −0.9% (test) below

[3]We compare against DG-MAML as the best *public* available model on the CSpider leaderboard at the time of writing.



	Cosine Sim. (↑)	Hausdorff (↓)
Train-EN∪All	0.47	6.48
TrainEN→FT-All	0.54	5.23
Reptile-EN→FT-All	0.67	4.47
XG-REPTILE	**0.85**	**2.46**

Figure 3: PCA Visualizations of sentence-averaged encodings for English (EN), French (FR), and Chinese (ZH) from the ATIS test set (@1% sampling from Table 2). We identify the regularized weight manifold that improves cross-lingual transfer using XG-REPTILE. We also improve in two similarity metrics averaged across languages.

the monolingual DG-MAML model. While performance does not match monolingual models, the multilingual approach has additional utility in serving more users. As a zero-shot setup, predicting SQL from CSpider inputs through the model trained for English yields 7.9% validation accuracy, underscoring that cross-lingual transfer for this dataset is non-trivial.

Varying the target sample size demonstrates more variable effects for Spider compared to ATIS. Notably, increasing the sample size yields poorer English performance beyond the optimal XG-REPTILE@5% setting for English. This may be a consequence of the cross-database challenge in Spider—information shared across languages may be less beneficial than for single-domain ATIS. The least performant model for both languages is XG-REPTILE@1%. Low performance here for Chinese can be expected, but the performance for English is surprising. We suggest that this result is a consequence of ''over-sampling'' of the target data disrupting the overall training process. That is, for 1% sampling and optimal $K = 4$, the target data is ''over-used'' $25\times$ for each epoch of support data. We further observe diminishing benefits for English with additional Chinese samples. While we trained a competitive parser with minimal Chinese data, this effect could be a consequence of how RAT-SQL cannot exploit certain English-oriented learning features (e.g., lexical similarity scores). Future work could explore cross-lingual strategies to unify entity modeling for improved feature sharing.

Visualizing the Manifold Analysis of XG-REPTILE in Section 4 relies on a theoretical basis that first-order meta-learning creates a dense high-likelihood sub-region in the parameters (i.e., *optimal manifold*). Under these conditions, representations of new domains should cluster within the manifold to allow for rapid adaptation with minimal samples. This contrasts with methods without meta-learning, which provide no guarantees of representation density. However, metrics in Tables 2 and 3 do not directly explain if this expected effect arises. To this end, we visualize ATIS test set encoder outputs using PCA (Halko et al., 2011) in Figure 3. We contrast English (support) and French and Chinese as the most and least similar target languages. Using PCA allows for direct interpretation of low-dimensional distances across approaches. Cross-lingual similarity is a proxy for manifold alignment—as our goal is accurate cross-lingual transfer from

closely aligned representations from source and target languages (Xia et al., 2021; Sherborne and Lapata, 2022).

Analyzing Figure 3, we observe meta-learning methods (*Reptile-EN→FT-All*, XG-REPTILE) to fit target languages closer to the support (English, yellow circle). In contrast, methods not utilizing meta-learning (*Train-EN∪All*, *Train-EN→FT-All*) appear less ordered with weaker representation overlap. Encodings from XG-REPTILE are less separable across languages and densely clustered, suggesting the regularized manifold hypothesized in Section 4 ultimately yields improved cross-lingual transfer. Visualizing encodings from English in the *Reptile-EN* model *before* fine-tuning produces a similar cluster (not shown), however, required fine-tuning results in ‘‘spreading’’ leading to less cross-lingual similarity.

We also quantitatively examine the average encoding change in Figure 3 using cosine similarity and Hausdorff distance (Patra et al., 2019) between English and each target language. Cosine similarity is measured pair-wise across parallel inputs in each language to gauge similarity from representations with equivalent SQL outputs. As a measure of mutual proximity between sets, Hausdorff distance denotes a worst-case distance between languages to measure more general ‘‘closeness’’. Under both metrics, XG-REPTILE yields the best performance with the most substantial pair-wise similarity and Hausdorff similarity. These indicators for cross-lingual similarity further support the observation that our expected behavior is legitimately occurring during training.

Our findings better explain *why* our XG-REPTILE performs above other training algorithms. Specifically, our results suggest that XG-REPTILE learns a *regularized manifold* which produces stronger cross-lingual similarity and improved parsing compared to Reptile *fine-tuning a manifold*. This contrast will inform future work for cross-lingual meta-learning where XG-REPTILE can be applied.

Error Analysis We can also examine *where* the improved cross-lingual transfer influences parsing performance. Similar to Figure 3, we consider the results of models using 1% sampling as the worst-case performance and examine where XG-REPTILE improves on other methods on the test set (448 examples) over five languages.

EN	Show me all flights from San Jose to Phoenix
FR	Me montrer tous les vols de San Josè á Phoenix
×	SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1, airport_service airport_service_2, city city_2 WHERE flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'SAN FRANCISCO' AND flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'PHILADELPHIA' ;
✓	SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1, airport_service airport_service_2, city city_2 WHERE flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'SAN JOSE' AND flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'PHOENIX' ;

Figure 4: Contrast between SQL from a French input from ATIS for *Train-EN∪All* and XG-REPTILE. The entities ‘‘San José’’ and ‘‘Phoenix’’ are not observed in the 1% sample of French data but are mentioned in the English support data. The *Train-EN∪All* approach fails to connect attributes seen in English when generating SQL from French inputs (×). Training with XG-REPTILE better leverages support data to generate accurate SQL from other languages (✓).

Accurate semantic parsing requires sophisticated entity handling to translate mentioned proper nouns from utterance to logical form. In our few-shot sampling scenario, *most* entities will appear in the English support data (e.g., ‘‘Denver’’ or ‘‘American Airlines’’), and *some* will be mentioned within the target language sample (e.g., ‘‘Mineápolis’’ or ‘‘Nueva York’’ in Spanish). These samples cannot include all possible entities—effective cross-lingual learning must ‘‘connect’’ these entities from the support data to the target language—such that these names can be parsed when predicting SQL from the target language. As shown in Figure 4, the failure to recognize entities from support data, for inference on target languages, is a critical failing of all models besides XG-REPTILE.

The improvement in cross-lingual similarity using XG-REPTILE expresses a specific improvement in entity recognition. Compared to the worst performing model, *Train-EN∪All*, 55% of improvement accounts for handling entities absent from the 1% target sample but present in the 99% English support data. While XG-REPTILE can generate accurate SQL, other models are limited in expressivity to fall back on using seen entities from the 1% sample. This notably accounts for 60% of improvement in parsing Chinese, with minimal orthographic overlap to English, indicating that XG-REPTILE better leverages support data without reliance on token similarity. In 48% of improved parses, entity mishandling is the *sole error*—highlighting how limiting poor cross-lingual transfer is for our task.

Our model also improves handling of novel *modifiers* (e.g., ''on a weekday'', ''round-trip'') absent from target language samples. Modifiers are often realized as additional sub-queries and filtering logic in SQL outputs. Comparing XG-REPTILE to *Train-EN∪All*, 33% of improvement is related to modifier handling. Less capable systems fall back on modifiers observed from the target sample or ignore them entirely to generate inaccurate SQL.

While XG-REPTILE better links parsing knowledge from English to target languages—the problem is not solved. Outstanding errors in all languages primarily relate to query complexity, and the cross-lingual transfer gap is not closed. Furthermore, our error analysis suggests a future direction for optimal sample selection to minimize the error from interpreting unseen phenomena.

7 Conclusion

We propose XG-REPTILE, a meta-learning algorithm for few-shot cross-lingual generalization in semantic parsing. XG-REPTILE is able to better utilize fewer samples to learn an economical multilingual semantic parser with minimal cost and improved sample efficiency. Compared to adjacent training algorithms and zero-shot approaches, we obtain more accurate and consistent logical forms across languages similar and dissimilar to English. Results on ATIS show clear benefit across many languages and results on Spider demonstrate that XG-REPTILE is effective in a challenging cross-lingual and cross-database scenario. We focus our study on semantic parsing, however, this algorithm could be beneficial in other low-resource cross-lingual tasks. In future work we plan to examine how to better align entities in low-resource languages to further improve parsing accuracy.

Acknowledgments

We thank the action editor and anonymous reviewers for their constructive feedback. The authors also thank Nikita Moghe, Seraphina Goldfarb-Tarrant, Ondrej Bohdal, and Heather Lent for their insightful comments on earlier versions of this paper. We gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (grants EP/L016427/1 (Sherborne) and EP/W002876/1 (Lapata)) and the European Research Council (award 681760, Lapata).

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1253

Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta. 2020. Low-resource domain adaptation for compositional task-oriented semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5090–5100, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1269

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 379–389, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-1038

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML* 2017, Sydney, NSW, Australia, 6–11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Ofer Givoli and Roi Reichart. 2019. Zero-shot semantic parsing for instructions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4454–4464, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1438

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2019. Coupling retrieval and meta-learning for context-dependent semantic parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 855–866, Florence, Italy. Association for Computational Linguistics.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2011. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288. https://doi.org/10.1137/090771806

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.201

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*. https://doi.org/10.3115/116580.116613

Jonathan Herzig and Jonathan Berant. 2017. Neural semantic parsing over multiple knowledge-bases. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 623–628, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2098

Jonathan Herzig and Jonathan Berant. 2021. Span-based semantic parsing for compositional generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 908–921, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.74

Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. 2022. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(09):5149–5169.

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 the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Po-Sen Huang, Chenglong Wang, Rishabh Singh, Wen-tau Yih, and Xiaodong He. 2018. Natural language to structured query generation via meta-learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 732–738, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2115

Zhanming Jie and Wei Lu. 2014. Multilingual semantic parsing: Parsing multiple languages into semantic representations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1291–1301, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Akhil Kedia, Sai Chetan Chinthakindi, and Wonho Ryu. 2021. Beyond Reptile: Meta-learned dot-product maximization between gradients for improved single-task regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 407–420, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.37

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2012

Dongjun Lee, Jaesik Yoon, Jongyun Song, Sang-gil Lee, and Sungroh Yoon. 2019. One-shot learning for text-to-sql generation. *CoRR*, abs/1905.11499.

Hung-yi Lee, Ngoc Thang Vu, and Shang-Wen Li. 2021. Meta learning and its applications to natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 15–20, Online. Association for Computational Linguistics.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). https://doi.org/10.1609/aaai.v32i1.11596

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 the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal,

Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.484

Qingkai Min, Yuefeng Shi, and Yue Zhang. 2019. A pilot study for Chinese SQL semantic parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3652–3658, Hong Kong, China. Association for Computational Linguistics.

Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. 2020. Localizing open-ontology QA semantic parsers in a day using machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983, Online. Association for Computational Linguistics.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999v3.

Massimo Nicosia, Zhongdi Qu, and Yasemin Altun. 2021. Translate & fill: Improving zero-shot multilingual semantic parsing with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3272–3284, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.279

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.368

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1018

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Kyle Richardson, Jonathan Berant, and Jonas Kuhn. 2018. Polyglot semantic parsing in APIs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 720–730, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1066

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.779

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.285

Tom Sherborne, Yumo Xu, and Mirella Lapata. 2020. Bootstrapping a crosslingual semantic

parser. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 499–517, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.45

Yu Su and Xifeng Yan. 2017. Cross-domain semantic parsing via paraphrasing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1235–1246, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1127

Alane Suhr, Ming-Wei Chang, Peter Shaw, and Kenton Lee. 2020. Exploring unexplored generalization challenges for cross-database semantic parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8372–8388, Online. Association for Computational Linguistics.

Yibo Sun, Duyu Tang, Nan Duan, Yeyun Gong, Xiaocheng Feng, Bing Qin, and Daxin Jiang. 2020. Neural semantic parsing in low-resource settings with back-translation and meta-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8960–8967. https://doi.org/10.1609/aaai.v34i05.6427

Raymond Hendy Susanto and Wei Lu. 2017a. Neural architectures for multilingual semantic parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–44, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2007

Raymond Hendy Susanto and Wei Lu. 2017b. Semantic parsing with neural hybrid trees. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.304

Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2018. (Almost) Zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP* 2018, Calgary, AB, Canada, April 15-20, 2018, pages 6034–6038. IEEE.

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 the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.197

Bailin Wang, Mirella Lapata, and Ivan Titov. 2021a. Meta-learning for domain generalization in semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–379, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.33

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020a. RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7567–7578, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.677

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. 2021b. Generalizing to unseen domains: A survey on domain generalization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4627–4635. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2021/628

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020b. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Survey*, 53(3). https://doi.org/10.1145/3386252

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144v2.

Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. 2021. MetaXL: Meta representation transformation for low-resource cross-lingual learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 499–511, Online. Association for Computational Linguistics.

Silei Xu, Sina Semnani, Giovanni Campagna, and Monica Lam. 2020a. AutoQA: From databases to QA semantic parsers with only synthetic training data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 422–434, Online. Association for Computational Linguistics.

Weijia Xu, Batool Haider, Jason Krone, and Saab Mansour. 2021. Soft layer selection with meta-learning for zero-shot cross-lingual transfer. In *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*, pages 11–18, Online. Association for Computational Linguistics.

Weijia Xu, Batool Haider, and Saab Mansour. 2020b. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Jingfeng Yang, Federico Fancellu, Bonnie Webber, and Diyi Yang. 2021. Frustratingly simple but surprisingly strong: Using language-independent features for zero-shot cross-lingual semantic parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5848–5856, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.472

Pengcheng Yin and Graham Neubig. 2017. A syntactic neural model for general-purpose code generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 440–450, Vancouver, Canada. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1425

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.447

Victor Zhong, Mike Lewis, Sida I. Wang, and Luke Zettlemoyer. 2020. Grounded adaptation for zero-shot executable semantic parsing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6869–6882, Online. Association for Computational Linguistics.

Qile Zhu, Haidar Khan, Saleh Soltan, Stephen Rawls, and Wael Hamza. 2020. Don't parse, insert: Multilingual semantic parsing with insertion based decoding. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 496–506, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.conll-1.40