Semantic adapters in text-to-SQL for low-resource languages: the importance of semantic information

Anton Bulle Labate
Universidade de São Paulo
São Paulo, Brazil
antonlabate@usp.br

Fabio Gagliardi Cozman
Universidade de São Paulo
São Paulo, Brazil
fgcozman@usp.br

Abstract

This paper investigates whether injecting semantic structural knowledge of low-resource or unfamiliar languages into Large Language Models (LLMs) enhances performance on downstream Text-to-SQL tasks. We evaluate our approach on Galician, a Romance low-resource language, and, to demonstrate its generality, also on Guarani, a (very) low-resource language of an entirely distinct linguistic profile. Our empirical results show that semantically-aware models consistently outperform baselines across all benchmark metrics.

1 Introduction

Large language models (LLMs) have shown proficiency in many languages by achieving remarkable results across a diverse range of multilingual benchmarks. Models such as gpt-oss (OpenAI et al., 2025)have achieved strong results in multilingual benchmarks, such as multilingual MMLU (Hendrycks et al., 2021) and Multilingual Grade School Math (MGSM) (Shi et al., 2023), closing the gap between their proprietary counter parts. However, large language models still struggle with low-resource languages (Pava et al., 2025).

Recent work has proposed techniques for LLM development in low-resource languages, including few-shot in-context learning with semantically aligned exemplars for tasks like translation and sentiment analysis (Cahyawijaya et al., 2024), parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) on models for languages such as Marathi (Khade et al., 2025). Despite these advances, the natural language–to–SQL task in low-resource languages remains largely unexplored.

There have been efforts in text-to-SQL (often referred to as text2sql) to languages beyond English. Jose and Cozman (2023) translated the Spider dataset (Yu et al., 2018), one of the benchmarks for the task, to cover Spanish, Portuguese and French, enabling the development of text-to-SQL models capable of operating directly on those languages. This eliminates the extra latency and token costs associated with English-centric pipelines, caused by the required additional intermediate step of translating the user queries into English prior to SQL generation. Min et al. (2019) manually translated the Spider dataset to Chinese. None of these efforts focused on a very low-resource language such as Galician.

The translation of natural language to SQL is a particularly interesting task, as it requires more of the model than only the overall understanding of the user query: it requires that model correctly understands how the terms within the sentence interact between themselves to convey the sentence's meaning, so that the model can translate these interactions to the final SQL query. This is already a harsh requirement for models using English as input language, as LLMs have already been shown to have difficulties with tasks involving sentence compositionality (Leivada et al., 2023). For languages to which the model has been less exposed during pre-training, and hence is less familiar with its sentences' dependency and semantic structures, this issue becomes particularly acute.

Prior work across a diverse range of tasks has demonstrated that incorporating sentence-structure information can be beneficial for models (Labate and Cozman, 2024; Currey and Heafield, 2019; Qian et al., 2021). For example, Currey and Heafield (2019) enhance translation performance by infusing morphology trees as input to trained encoders, and also improve translation with encoder-decoder models by pre-training the model on a parsing task.

In this paper, we introduce a method that exploits semantic information to enhance a model's understanding of sentence structure and meaning in low-resource or unseen languages, without requiring extensive pre-training on those languages.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1032–1037

We evaluate our method in the text-to-SQL task, as it is highly sensitive to accurate semantic interpretation and to the relationships between sentence constituents, and therefore particularly well suited to benefit from semantic enhancement. Our method follows a two-stage adapter-based procedure. First, we train an adapter on a semantic parsing task and then merge it with the base language model, yielding a semantically aware model. We next use this resulting model as the foundation for fine-tuning a second adapter for the text-to-SQL task[1].

We conduct experiments on Galician, a Romance language that, despite its proximity to Portuguese, remains a low-resource language. This setting allows us to examine whether semantic enhancement provides additional benefits beyond those potentially obtained through cross-lingual sentence-level knowledge transfer from related high-resource languages. To further assess the generality of our approach for languages largely unfamiliar to current language models, we also evaluate it on Guarani, a very low-resource language. Across both languages, our results show that the semantically enhanced model consistently outperforms a baseline model whose adapter is fine-tuned directly for text-to-SQL. In addition, we have generated text-to-SQL models for Galician and Guarani, which, to the best of our knowledge, are the first of their kind.

2 Semantic adapters in text-to-SQL

Extensive pre-training on high-resource languages enables language models to internalize linguistic structure in an unsupervised manner, giving rise to emergent abilities such as syntactic–knowledge alignment (Hewitt and Manning, 2019). In fact, Futrell et al. (2019) state that the representation of syntactic structure requires either syntactic supervision or very large datasets.

Building on prior work demonstrating that explicitly incorporating linguistic structure can benefit low-resource settings (Deng and Wang, 2024), we set to use sentence-level semantic information to increase the model's familiarity with sentence structure and, consequently, with the interactions among core terms that determine meaning. More precisely, drawing further inspiration from approaches that incorporate syntactic information through joint training on downstream tasks and syntactic parsing (Qian et al., 2021; Currey and Heafield, 2019), we integrate semantic information into our model in a two-stage adapter-based approach, as follows.

First, we fine-tune an adapter on top of a base language model to generate semantic trees for the input sentences. After training, this adapter is merged with the base model to produce a semantically-aware language model. In the second stage, we fine-tune a second adapter on this semantically-aware model for our downstream task, text-to-SQL, using data in the same low-resource language for which semantic parsing was performed.

Figure 1 illustrates the resulting end-to-end text-to-SQL inference pipeline with a Galician input sentence. Notably, in contrast to prior work, our approach does not require users to perform semantic parsing or provide semantic annotations at inference time; instead, the model internalizes semantic knowledge during training, eliminating the need for explicit semantic inputs.

We focus on semantic information rather than syntactic annotations because semantic representations explicitly capture the relationships among meaning-bearing elements of a sentence, while abstracting away function words that do not contribute directly to its core meaning (e.g., determiners). As a result, we expose the model to a concise representation of how meaning is composed through relations among key terms, while keeping the input representation lean and mitigating known issues in large language models such as the "needle-in-a-haystack" and "lost-in-the-middle" effects that arise with long inputs.

3 Experiments and discussion

This section evaluates the effectiveness of our semantic-enhancement approach for low-resource languages. We evaluate our approach on the text-to-SQL task, a benchmark that is highly sensitive to accurate semantic interpretation and compositional reasoning, and examine whether pre-exposing the model to sentence-level semantic structure improves downstream performance. Experiments are run with two low-resource languages, Galician and Guarani. We compare a semantically enhanced language model with a baseline that directly fine-tunes adapters for text-to-SQL, keeping model architecture and training conditions fixed.

To obtain the semantic adapter, which is later merged with the base language model to produce the semantically-aware LM, we train an adapter

[1]The fine-tuning code is available at https://github.com/antonlabate/semantic-infusion

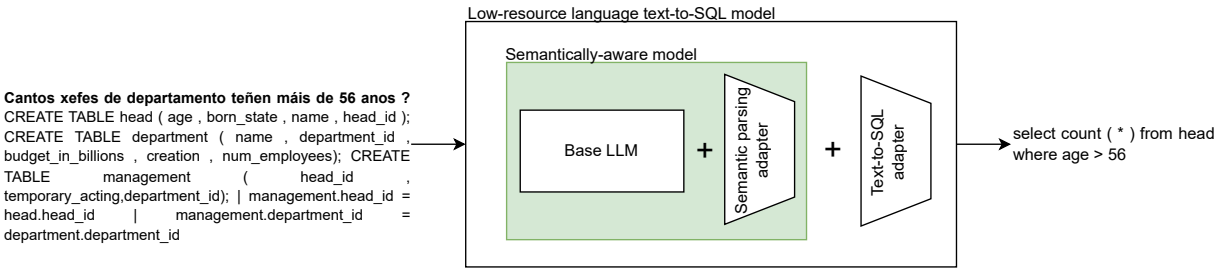

Figure 1: The end-to-end inference pipeline, depicting both semantic and text-to-SQL trained adapters.

on a semantic parsing task. As a representation of sentence-level semantic information, we adopt Abstract Meaning Representation (AMR), one of the most influential and widely used semantic formalisms in the field. However, to the best of our knowledge, no publicly available AMR corpora exist for Galician or Guarani. Consequently, we construct our own AMR datasets for these languages by translating English sentences and projecting their corresponding AMR annotations.

AMR is designed to capture sentence meaning independently of surface realization. Although languages differ substantially in their syntactic and morphological realizations, sentences that convey the same intent typically share the same underlying semantic concepts—and, consequently, similar, if not the same, AMR structures. As a result, AMR provides a normalized representation that encodes what is meant rather than how it is expressed. Since the Galician and Guarani sentences are translations of the original English sentences and preserve their propositional content, we assume that the underlying AMR graphs remain valid across languages. This assumption is consistent with prior work on cross-lingual and multilingual AMR parsing (Damonte and Cohen, 2018).

To expose the model to the basic semantic structures of Galician and Guarani using sentences aligned with the text-to-SQL domain, we select a subset of the high-quality synthetic text-to-SQL dataset introduced by Meyer et al. (2024). This dataset spans a wide range of domains and SQL operations and has been validated by its authors. Because our goal is to teach the model the basics of sentence-level semantic structure in Galician and Guarani—rather than to develop a specialized semantic parsing system capable of modeling fine-grained semantic distinctions—we focus on instances with relatively simple sentence structures. Accordingly, we restrict our training data to examples categorized as "basic SQL" in the original dataset. This results in a training dataset with 46,042 instances, and test set comprising 2,424 examples.

The AMR graphs for the English sentences were generated using amrlib[2], using a BART large as base model for the graph generation. The English sentences were translated to Galician and Guarani using the NLLB-200-3.3B model (Costa-jussà et al., 2022). The NLLB model is a well-known machine translation model for addressing the translation to 200 languages, including low-resource ones, with state-of-the-art results. The accuracy of this model is evaluated in the FLORES-200 benchmark and its results for the languages can be found in the original paper, which aims to support the preservation of such languages by enabling translation to and from English.[3].

The text-to-SQL adapters were trained and evaluated using translated versions of the Spider's (Yu et al., 2018) training and development sets, a popular benchmark in the text-to-SQL literature. These splits are comprised of 7000 and 1034 examples each, respectively. Again, the original English sentences from the Spider dataset were translated using the NLLB-200-3.3B model.[4]

We acknowledge that neural machine translation (NMT) can incur in translation errors and is not as reliable a human translation, occasionally distorting particular and contextual expressions in the source language. Nevertheless, in low-resource settings the available alternatives are lim-

[2]https://github.com/bjascob/amrlib

[3]The generated dataset for fine-tuning the Galician and Guarani semantic adapters can be found at https://huggingface.co/datasets/antonlabate/synthetic-simple-text-to-AMR-SQL-pt-glg-grn-cym

[4]The translated Spider splits to Galician and Guarani are available at https://huggingface.co/datasets/antonlabate/spider-glg-grn

ited, and we therefore rely on the NLLB model, a well-established and extensively validated machine translation system. We expect that higher-quality translations would likely lead to further performance gains. However, the primary objective of this work is not to maximize absolute text-to-SQL accuracy, but to assess whether language models benefit from exposure to semantic information when performing the text-to-SQL task. Consequently, our analysis focuses on the relative performance difference between a semantically aware model fine-tuned with a text-to-SQL adapter and a baseline model in which the adapter is fine-tuned directly on the base language model under identical conditions. Performance is evaluated using Exact Set Match (ESM) and Execution Accuracy (EXA).

For both languages, the fine-tuning settings were the same. The base model used was Mistral-7B-Instruct-v0.3. Because the Spider development set translated into Galician and Guarani was created by us and, to the best of our knowledge, no public versions are available, the risk of data contamination is very unlikely. Adapter fine-tuning is performed using QLoRA, with model weights loaded in 4-bit NF4 precision. The adapters use a rank of 32, a LoRA scaling factor (α) of 64, and a dropout rate of 0.05. Supervised fine-tuning is conducted with a learning rate of 2e-5, a warmup ratio of 0.03, and a constant learning-rate scheduler. Each adapter is fine-tuned for one epoch at each stage (semantic parsing and text-to-SQL), using an effective batch size of 16. We employ the paged AdamW optimizer in 8-bit precision. All experiments are conducted on a computing infrastructure equipped with two NVIDIA GeForce RTX 3090 GPUs with 24 GB of memory each.

Table 1 presents a comparison between the semantically aware model fine-tuned with a text-to-SQL adapter and the baseline model fine-tuned with a text-to-SQL adapter, evaluated using Exact Set Match (ESM) and Execution Accuracy (EXA). The table reports the scores achieved by each model on the text-to-SQL task under both metrics. In addition, it includes the relative performance differences between the baseline (base model with a text-to-SQL adapter) and the semantically aware model, denoted as ΔESM and ΔEXA.

The results in Table 1 demonstrate that models integrated with semantic adapters achieved the highest overall performance on the text2sql task for both target languages. Specifically, the configuration utilizing text2sql adapters trained atop

Galician	**ESM**	**EXA**	**ΔESM**	**ΔEXA**
Base model	0.647	0.658	-	-
Semantic model	0.656	0.665	1.39	1.06
Guarani	**ESM**	**EXA**	**ΔESM**	**ΔEXA**
Base model	0.413	0.421	-	-
Semantic model	0.480	0.487	16.22	15.68

Table 1: Text-to-SQL performance comparison for Galician and Guarani between semantically aware language models with text-to-SQL adapters and baseline language models with directly fine-tuned text-to-SQL adapters, evaluated under ESM and EXA. Relative performance differences are reported in percentages.

semantically-merged models yielded the highest scores across both ESM and EXA metrics, outperforming their corresponding baselines. We observe that the performance boost was more pronounced for Guarani than for Galician. This disparity between Galician and Guarani likely arises from the linguistic proximity between Galician and Portuguese; since the base model is likely already familiar with the latter, leveraging its existing knowledge of Portuguese sentences to process the linguistically similar Galician, the semantic adapter provides less useful novel information for Galician than it does for the lower-resource Guarani. This suggests that the semantic enhancement is especially beneficial for languages with which the base model has minimal prior exposure. Furthermore, while the semantic-enhanced Galician model excelled in easy-to-medium queries, its performance degraded on complex queries. This performance drop likely stems from the semantic adapter's training data, which was restricted to "easy" query structures. Consequently, when presented with harder queries, lacking sufficient exposure to the corresponding semantic patterns, the model may be less effective at decomposing and interpreting their structure and intent compared to simpler cases.

4 Conclusion

In this paper, we investigated whether exposing LLMs to semantic knowledge in previously unfamiliar languages can improve performance on the downstream text-to-SQL task. We evaluate our approach on Galician and Guarani, two low-resource languages. For both languages, semantically aware models achieve the highest scores under the evaluated metrics, indicating that training on semantic parsing provides useful linguistic inductive biases for text-to-SQL. The gains are particularly pro-

nounced for Guarani, a language with which the base model is likely to have minimal prior exposure, aligning with our goal of improving structural awareness in previously unfamiliar languages. In addition, as contributions of this work, we generated the first text-to-SQL models and semantic parsers for Galician and Guarani to the best of our knowledge, along with the release of the datasets used to train them.

Acknowledgements

The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and from the IBM Corporation. F. G. C. was partially supported by CNPq grants 312180/2018-7 and 305753/2022-3. The authors also thank support by CAPES - Finance Code 001.

References

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Anna Currey and Kenneth Heafield. 2019. Incorporating source syntax into transformer-based neural machine translation. In *Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 24–33, Florence, Italy. Association for Computational Linguistics.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual Abstract Meaning Representation parsing. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, Louisiana. Association for Computational Linguistics.

DeLin Deng and LiQing Wang. 2024. Improving low-resource machine translation using syntactic dependencies. In *2024 6th World Symposium on Artificial Intelligence (WSAI)*, pages 29–33.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 32–42, Minnesota. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minnesota. Association for Computational Linguistics.

Marcelo Archanjo Jose and Fabio Gagliardi Cozman. 2023. A multilingual translator to SQL with database schema pruning to improve self-attention. *International Journal of Information Technology*, 15(6):3015–3023.

Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar, and Raviraj Joshi. 2025. Challenges in adapting multilingual LLMs to low-resource languages using LoRA PEFT tuning. In *First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 217–222, Abu Dhabi. International Committee on Computational Linguistics.

Anton Bulle Labate and Fabio Gagliardi Cozman. 2024. Infusing prompts with syntax and semantics. *Preprint*, arXiv:2412.06107.

Evelina Leivada, Elliot Murphy, and Gary Marcus. 2023. Dall·e 2 fails to reliably capture common syntactic processes. *Social Sciences Humanities Open*, 8(1):100648.

Yev Meyer, Marjan Emadi, Dhruv Nathawani, Lipika Ramaswamy, Kendrick Boyd, Maarten Van Segbroeck, Matthew Grossman, Piotr Mlocek, and Drew Newberry. 2024. Synthetic-Text-To-SQL: A synthetic dataset for training language models to generate sql queries from natural language prompts.

Qingkai Min, Yuefeng Shi, and Yue Zhang. 2019. A pilot study for Chinese SQL semantic parsing. In *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.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien

Bubeck, and 108 others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Juan Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T Truong, Daniel Zhang, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the (language) gap: Mapping the challenges of llm development in low-resource language contexts.

Peng Qian, Tahira Naseem, Roger Levy, and Ramón Fernandez Astudillo. 2021. Structural guidance for transformer language models. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3735–3745, Online. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations*.

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 *Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.