VLSP 2025 challenge: Vietnamese Semantic Parsing

Ha My Linh, Pham Thi Duc, Le Ngoc Toan, Nguyen Thi Minh Huyen
VNU University of Science, Hanoi Vietnam
{hamylinh, phamthiduc, lengoctoan_t65, huyenntm}@hus.edu.vn
Correspondence: hamylinh@hus.edu.vn

Abstract

In 2025, the Eleven Workshop on Vietnamese Language and Speech Processing (VLSP 2025) organized the first shared task on Vietnamese Semantic Parsing (viSemParse). The primary objective of the viSemParse challenge is to evaluate the capability of systems in capturing the underlying semantic structures of Vietnamese sentences. The gold-standard datasets developed specifically for this task serve as essential resources for training and evaluating semantic parsers.

The viSemParse 2025 dataset comprises 2,500 sentences, divided into three subsets: training, public test, and private test. The competition was conducted on the AIHub platform, where systems were ranked based on their performance in the private test phase following a public testing period. The best-performing system in the viSemParse - VLSP 2025 shared task achieved a Smatch score of 58%, highlighting both the challenges and the potential for advancement in Vietnamese semantic parsing research.

Keywords: Vietnamese semantic parsing, viSemParse, VLSP 2025

1 Introduction

Semantic parsing is a core task in natural language understanding that converts sentences into structured representations capturing their underlying meaning. Recently, representation models and semantically annotated corpora have been actively developed to formalize the meaning of words, sentences, and texts. Semantic representation models facilitate understanding and interpreting language across different contexts, addressing issues of ambiguity and semantic vagueness. Notable annotated corpora include PropBank (Kingsbury and Palmer, 2002), which provides shallow role-based annotations, AMR (Banarescu et al., 2013), which offers deep annotations using an abstract meaning representation framework, as well as other semantic resources such as The Groningen Bank (GMB) (Bos, 2013), Universal Conceptual Cognitive Annotation (UCCA) (O. and Rappoport, 2013), and Uniform Meaning Representation (UMR) (Chun and Xue, 2024), ...

Among existing semantic formalisms, Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is widely used due to its ability to represent predicate–argument structures, core semantic roles, and concept-to-concept relations. AMR parsing has made significant progress in English, largely driven by neural sequence-to-sequence architectures and large language models (LLMs). Notable examples include SPRING (Bevilacqua et al., 2021), based on the T5 model (Raffel et al., 2020), and AMRBART (Cai and Lam, 2020), which enhances BART with graph-aware pretraining. These approaches achieve state-of-the-art performance, enabling accurate generation of graph-based meaning representations.

For Vietnamese, several studies have explored the adaptation of AMR for semantic parsing. Linh et al. (Linh and Nguyen, 2019) proposed modifications to the AMR framework to account for syntactic differences between English and Vietnamese, laying the groundwork for annotating Vietnamese sentences with AMR. More recently, Regan et al. (Regan et al., 2024) introduced MASSIVE-AMR, a dataset containing over 84,000 text-to-graph annotations, currently the largest and most diverse of its kind, with AMR graphs for 1,685 information-seeking utterances mapped to more than 50 languages, including Vietnamese. Building a high-quality semantic role-labeled corpus and developing semantic analysis tools is crucial for low-resource languages like Vietnamese, aiming to promote the growth of the Vietnamese natural language processing research and application community.

The VLSP 2025 Shared Task on Vietnamese Semantic Parsing[1] (viSemParse) was organized as the first large-scale benchmark for AMR-style semantic parsing in Vietnamese. The shared task provides a manually annotated corpus of 2,500 sentences in PENMAN format (Goodman, 2020), divided into training, public test, and private test sets. The primary goal is to evaluate system performance in capturing semantic relations and structural consistency while encouraging research on multilingual transfer learning and graph generation for low-resource languages. Participants were invited to develop models that could automatically generate semantic graphs from Vietnamese sentences and submit predictions for evaluation using the Smatch metric (Cai and Knight, 2013).

Participating teams adopted a wide variety of modeling strategies. Several systems utilized sequence-to-sequence transformer architectures fine-tuned directly on the viSemParse dataset to learn the mapping between sentences and linearized AMR graphs. Others explored fine-tuning Large Language Models (e.g., Qwen3-14B (Team, 2025), Gemma-3 (DeepMind, 2025), and Phi-4 (Abdin et al., 2025)) through instruction-based methods or low-rank adaptation (LoRA) (Hu et al., 2022) to improve training efficiency. In addition, specialized frameworks integrating data normalization, variable restoration, and post-processing rules were developed to enhance structural consistency and validity. Overall, these diverse approaches mark a meaningful advancement in Vietnamese semantic representation, providing valuable insights and establishing a strong foundation for future research on low-resource semantic parsing.

In this paper, we present a comprehensive overview of viSemParse Shared Task, providing a detailed description of the task definition, dataset construction, participating systems, and evaluation framework. We also discuss the significance of this initiative in advancing Vietnamese semantic parsing research and its broader implications for developing language understanding applications in low-resource settings. This shared task represents the first large-scale effort toward building and evaluating models capable of generating AMR-style semantic representations for Vietnamese, addressing the growing demand for deeper semantic understanding and cross-lingual transfer learning in Vietnamese NLP.

2 Shared task description

The objective of developing a Vietnamese semantic parser is to achieve accurate understanding and formal representation of Vietnamese sentences by analyzing their syntactic and semantic structures. The parser is designed to extract the underlying meaning of text and convert it into structured representations, such as Abstract Meaning Representation (AMR) or logical forms. This shared task focuses on creating a Vietnamese benchmark dataset with semantic annotations and evaluating the performance of Vietnamese semantic parsing models.

A Vietnamese semantic parsing system operates as follows:

- *Input*: A natural sentence in Vietnamese.
- *Output*: The corresponding semantic representation in PENMAN format (Goodman, 2020), where nodes represent concepts and edges denote semantic relations between them.

Figure 1 illustrates the semantic graph of the Vietnamese sentence “Anh nói rõ cho em nghe thử coi”. The main predicate “nói” (say) serves as the root node, with “Anh” (you) as the agent (:agent) and “rõ” (clearly) as the manner (:manner). The purpose relation (:purpose) connects this event to the secondary action “em nghe thử coi” (listen). Overall, the graph abstracts the syntactic surface into a structured meaning representation that highlights the relations between actions, participants, and intent.

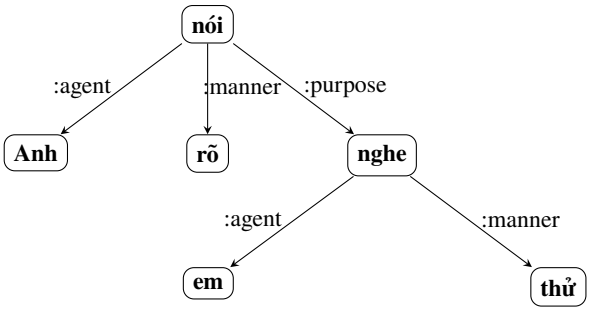

Figure 1: Semantic representation for the sentence “*Anh nói rõ cho em nghe thử coi.*”

The Smatch metric is widely used to evaluate the performance of semantic parsing systems. Smatch[2]

[1] https://vlsp.org.vn/vlsp2025/eval/visemparse

[2] https://github.com/snowblink14/smatch

(Cai and Knight, 2013) is an evaluation tool for AMR. It measures the similarity between two AMR graphs by finding a mapping between variables (nodes) that maximizes the number of matching triples (edges), denoted as M.

- M: the number of matching triples between two AMR graphs
- T: the total number of triples in the first AMR graph (Predicted AMR graph)
- G: the total number of triples in the second AMR graph (Gold AMR graph)

The precision, recall, and Smatch score (F_1-score) are defined as follows:

$$P = \frac{M}{T} \quad \text{(Precision)}$$

$$R = \frac{M}{G} \quad \text{(Recall)}$$

$$F_1 = \frac{2PR}{P+R} \quad \text{(Smatch score)}$$

The Smatch score F_1 reflects the semantic similarity between two AMR graphs: the higher the score, the more semantically similar the graphs are.

For example, we have two sentences in PENMAN format below:

```
#::snt Anh nói rõ cho em nghe thử coi .
(n / nói
   :agent (a / Anh)
   :manner (r / rõ)
   :purpose (n1 / nghe
      :pivot (e / em)
      :manner (t / thử)))
```

(a) Gold AMR representation

```
#::snt Anh nói rõ cho em nghe thử coi .
(n / nói
   :agent (a / Anh)
   :manner (r / rõ)
   :purpose (n1 / nghe
      :pivot (e / em)
      :mod (t / thử)))
```

(b) System-generated AMR representation

Figure 2: Comparison of gold and system-generated semantic representations for the sentence "Anh nói rõ cho em nghe thử coi."

In this case:

Smatch-score = F_1-score = 0.92 (Precision: 0.92, Recall: 0.92).

The competition was hosted on AIHub[3] and consisted of two phases: public test and private test. Teams submitted their results and updated their best scores on the leaderboard. Final rankings were determined based on these results along with the submitted technical reports.

3 Data preparation

The construction of the Vietnamese Semantic Parsing dataset followed a systematic two-phase process designed to ensure both linguistic validity and annotation consistency. The goal was to develop a high-quality Vietnamese corpus aligned with the Abstract Meaning Representation (AMR) framework while accounting for the unique syntactic and semantic characteristics of the language.

3.1 Building semantic labels for Vietnamese

To develop the Vietnamese semantic label set, several differences between the ways meaning is expressed in English and Vietnamese were studied. Designing additional semantic labels to capture these components was considered essential. The goal of the semantic representation model is not only to answer the simple question "Who is doing what to whom," but also to include additional information such as where, when, why, and how. The main semantic roles in the Vietnamese representation model were combined from LIRICS (Petukhova and Bunt, 2008) and English AMR (Banarescu et al., 2013). Moreover, the label set was designed to address certain limitations of AMR by incorporating representations for co-reference, tense–aspect, and additional labels to express function words and modifiers.

The Vietnamese semantic labels consist of 29 core roles, 74 non-core roles, 18 labels related to time and location, and 5 sentence-type labels. The main labels in the Vietnamese semantic representation include:

- *Predicates*: which serve as the core of semantic structures and are typically expressed by verbs, adjectives, or nouns functioning as predicates.

[3] https://aihub.ml/competitions/951

- *Core roles*: derived and extended from English AMR (Banarescu et al., 2013) and LIRICS (Petukhova and Bunt, 2008), including 29 roles such as *agent*, *patient*, *theme*, *beneficiary*, *goal*, *time*, and *location*, etc.

- *Non-core roles*: comprising labels that capture Vietnamese-specific semantic phenomena such as *classifiers*, *compounding*, *tense*, *degree*, etc and various adjuncts for time, manner, or instrument.

  In addition, the framework includes representations for named entities, linked to Wikipedia using the :wiki attribute, and co-reference across sentences within a paragraph to support discourse-level semantics. Sentence types such as imperative, interrogative, and multi-sentence structures are also explicitly labeled, allowing the Vietnamese semantic framework to more accurately reflect the richness and diversity of the language.

Full definitions of the labels and detailed annotation guidelines are provided in the annotation manual[4].

3.2 Data Annotation

In the second phase, manual semantic annotation was carried out over a period of six months by a team of five linguistic experts. To ensure high annotation quality, each data package underwent two independent annotation rounds, followed by consistency checking and error reconciliation. This multi-round review process helped maintain both semantic correctness and cross-annotator agreement.

A dedicated web-based annotation tool was developed to streamline the labeling process, providing functions for sentence visualization, AMR graph editing, and automatic validation of variable references and relation types.

To ensure data quality, control and validation procedures were integrated into the workflow. Inter-annotator agreement (IAA) was measured using Smatch (Cai and Knight, 2013) on a representative subset, with discrepancies resolved through consensus discussions. Automated scripts also checked for variable mismatches, incomplete edge relations, and structural inconsistencies. The resulting corpus achieved high semantic and annotation consistency, with detailed agreement scores presented in Table 1.

Table 1: Agreement between five annotators.

Annotator	Annotator	Smatch
Anno1	Anno2	0.73
Anno2	Anno3	0.96
Anno4	Anno5	0.95
Anno5	Anno1	0.77
Anno3	Anno4	0.86
Average		**0.86**

3.3 Data statistics

The dataset was compiled from multiple trusted linguistic sources, including the VietTreebank (VTB) (Nguyen et al., 2009), The Little Prince, and the Vietnamese Dependency Parsing dataset (Linh et al., 2020). These sources collectively provided a diverse range of syntactic and semantic phenomena, ensuring that the final corpus covers a wide spectrum of real-world Vietnamese language structures. Table 2 presents detailed statistics of the dataset, which is divided into three subsets: training, public test, and private test.

Table 2: Statistics of the Vietnamese Semantic Parsing dataset.

Split	# Sentences	Avg. tokens	Avg. chars
Train	1,750	11.54	44.17
Public test	150	17.13	68.24
Private test	600	13.27	59.03

Figure 3 illustrates the frequency of selected labels in the corpus. Core semantic roles such as :mod, :agent, and :theme occur most frequently, highlighting their central role in representing "who does what to whom." Non-core roles, including :degree, :manner, :time, and :op arguments, also appear often, showing that the corpus captures additional contextual information such as manner, time, and degree. Overall, the label distribution demonstrates a balanced coverage of both core and supplementary roles, providing a solid foundation for training Vietnamese semantic parsers.

4 Vietnamese semantic parsing methods

The viSemParse 2025 shared task attracted participation from 12 teams, with a total of 338

[4] https://github.com/vietnamesedp/Thesis/tree/main/MeaningRepresentation/TaiLieu

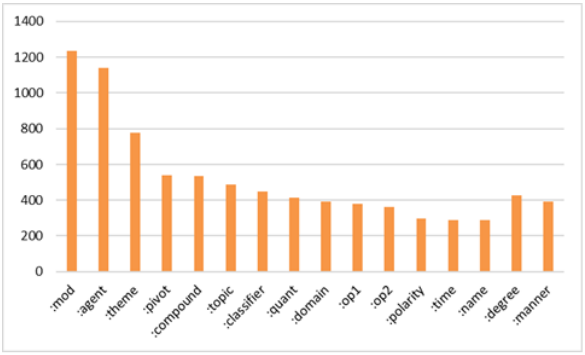

Figure 3: Frequency of semantic labels in Vietnamese dataset.

submissions across three phases: public test, private test. In this section, we provide an overview of the approaches adopted by the teams that submitted system description papers detailing their methods for Vietnamese semantic parsing.

4.1 Descriptions

Table 3 provides a summary of techniques used by the four best teams. Further details regarding the models are discussed in the following subsections.

Table 3: The models of the top four teams.

Team	Model	Fine-tuning	Optimization
UIT_BlackCoffee	Qwen3- 14B, Gemma-3, LLaMA-3.1, and Phi-4	LoRA fine-tuning	AdamW
ViAMR	Qwen3-1.7B	Supervised Fine-Tuning	AdamW
UIT-VNS	ViSemCrew - a multi-agent workflow	–	–
LangMind	BARTpho and ViT5	Fine-tuning	AdamW

Participating teams adopted diverse modeling approaches that combined both large-scale pre-trained language models and task-specific fine-tuning strategies. Several systems leveraged recent transformer-based LLMs (e.g., Qwen3, Gemma-3, LLaMA-3.1, Phi-4) with LoRA or supervised fine-tuning to adapt them to the Vietnamese semantic parsing task. Others explored custom architectures such as multi-agent workflows or Vietnamese-specific models like BARTpho and ViT5. Overall, these systems reflect a balanced integration of modern LLM adaptation techniques and linguistically informed designs, demonstrating the growing maturity of Vietnamese NLP research.

4.2 Participant approaches

Team UIT_BlackCoffee: The proposed method employs a three-stage pipeline for Vietnamese semantic parsing using large language models (LLMs), particularly Qwen3-14B (Team, 2025), Gemma-3 (DeepMind, 2025), Phi-4 (Abdin et al., 2025), and LLaMA-3.1 (Dubey et al., 2024), fine-tuned with Supervised Fine-Tuning (SFT) (Dong et al., 2023) with LoRA (Hu et al., 2022) for efficiency. The data preprocessing phase removes non-semantic elements such as variables and wiki tags, then linearizes AMR graphs into PENMAN-style text sequences to simplify training. After fine-tuning, a post-processing procedure restores variables and ensures structural validity by assigning unique identifiers to concept nodes, yielding well-formed AMR graphs. Using a 4-bit quantized Qwen3-14B model optimized with AdamW and early stopping, the approach effectively adapts multilingual LLMs to Vietnamese semantic parsing, improving both graph accuracy and syntactic consistency.

Team ViAMR: Adopts a three-stage framework combining preprocessing, supervised fine-tuning, and a constraint-aware inference pipeline for Vietnamese AMR parsing. In preprocessing, PENMAN-formatted strings are normalized into one-line sequences, missing brackets are fixed, and multiword nodes are converted with underscores to ensure syntactic validity. The model is then fine-tuned using Supervised Fine-Tuning (SFT) (Dong et al., 2023) on a compact decoder-only LLM backbone (Qwen3-1.7B (Team, 2025)), trained to map Vietnamese sentences to linearized AMR graphs under an instruction-following setup. Training employs AdamW with linear learning rate decay, gradient accumulation, and distributed optimization for efficiency. During inference, the system generates PENMAN strings and applies a series of string-level and graph-level repairs including role spacing, bracket balancing, variable deduplication, and canonicalization via PENMAN round-trip parsing to ensure well-formed, parsable outputs. This pipeline yields structurally consistent AMR graphs with stable evaluation performance while maintaining efficiency on modest computational resources.

Team UIT-VNS: The ViSemCrew framework tackles Vietnamese semantic parsing through

a multi-agent workflow that divides the task into specialized subtasks. It comprises five coordinated agents: the Linguistic Analysis Agent for morphological and syntactic processing, the Concept Extraction Agent for identifying semantic concepts, the Graph Construction Agent for building relations and determining the root predicate, the Validation Agent for ensuring graph correctness, and the Repair Agent for handling errors and regeneration. The agents operate sequentially with iterative validation and fallback mechanisms. Vietnamese - specific adaptations - such as passive voice handling, flexible word order, and implicit element recovery - along with a role reference database, further enhance parsing accuracy and consistency.

Team LangMind: The proposed method adopts an encoder–decoder transformer architecture for generating AMR-style semantic graphs from Vietnamese sentences. Two pretrained Vietnamese-specific models, BARTpho (Nguyen and Nguyen, 2022) and ViT5 (Nguyen et al., 2021), are fine-tuned using the official VLSP 2025 dataset. The data preprocessing pipeline includes sentence tokenization and normalization, depth-first traversal for graph linearization, and the removal of inconsistent AMRs to ensure high-quality input. The system explores tokenization strategies at both word and syllable levels, with punctuation treated as separate tokens to preserve structural integrity. During training, the models are optimized with AdamW (learning rate $5 * 10^{-5}$), batch size 8, and early stopping, using beam search (beam size $= 4$) for inference. The post-processing stage normalizes variable names, fixes bracket mismatches, and reconstructs AMR graphs from linearized sequences. Experiments reveal that BARTpho with word-level tokenization achieves the highest score, showing that fine-grained word segmentation contributes significantly to better graph connectivity and semantic accuracy in Vietnamese AMR parsing.

5 Results and discussion

In this section, we present the experimental results of the participating systems and provide a detailed discussion of their performance, common error patterns, and insights for future improvements.

5.1 Results

During the public test phase, we received 164 submissions, with the top-performing results summarized in Table 4. Team UIT_BlackCoffee achieved the highest Smatch score of 0.55, showing a strong balance between precision and recall. UIT-VNS ranked second with a score of 0.42, indicating competitive performance but still behind the leading team. ViAMR obtained a Smatch of 0.38, performing moderately, while LangMind scored 0.33, suggesting that their system requires significant improvements to match the performance of the top teams.

Table 4: Result of teams on the public test.

Rank	Team	P	R	Smatch (F_1)
1	UIT_BlackCoffee	0.53	0.57	0.55
2	UIT-VNS	0.40	0.45	0.42
3	ViAMR	0.35	0.41	0.38
4	LangMind	0.46	0.26	0.33

During the private test phase, a total of 174 submissions were received from all participating teams, with some changes in the rankings compared to the public test phase.

Table 5: Result of teams on the private test.

Rank	Team	P	R	Smatch (F_1)
1	UIT_BlackCoffee	0.52	0.64	0.58
2	ViAMR	0.42	0.50	0.46
3	UIT-VNS	0.40	0.44	0.42
4	LangMind	0.33	0.42	0.37

In Table 5, team UIT_BlackCoffee outperformed all other teams with a Smatch score of 0.58, demonstrating strong generalization to unseen data. ViAMR and UIT-VNS maintained relatively stable performance, scoring 0.46 and 0.42, respectively, while LangMind improved slightly to 0.37 but remained the lowest-ranking team. Overall, these results indicate that UIT_BlackCoffee's system generalizes most effectively, UIT-VNS performs well on familiar data but shows less robustness, and both LangMind and ViAMR require further improvements in precision and recall.

5.2 Errors Analysis

In this section, we provide a detailed error analysis on the private test set for the top-performing teams in the viSemParse shared task: UIT_BlackCoffee (Tables 6), ViAMR (Tables 7), UIT-VNS (Tables 8), and LangMind (Tables 9). The analysis highlights the most frequent label

mismatches, including incorrect predictions (*Pred*), missing labels, and extra labels.

Table 6: Error analysis on private test for Team UIT_BlackCoffee.

Gold	Pred	Count	Missing label	Count	Extra label	Count
pivot	agent	52	domain	92	agent	474
topic	theme	25	agent	57	op1	306
compound	direction	22	compound	49	domain	269
compound	manner	21	modality	36	op2	244
theme	topic	18	theme	32	theme	217
manner	degree	16	prep	32	ARG2	183
compound	theme	15	op1	30	ARG1	166
theme	patient	14	op2	30	name	132
theme	compound	12	manner	25	time	126
pivot	domain	12	topic	24	topic	120

Table 7: Error analysis on private test for Team ViAMR.

Gold	Pred	Count	Missing label	Count	Extra label	Count
pivot	agent	39	domain	113	agent	481
compound	manner	25	agent	88	op1	363
compound	direction	22	compound	74	domain	352
compound	theme	18	theme	64	op2	303
patient	theme	17	modality	62	theme	300
agent	theme	15	op1	61	ARG2	184
topic	theme	14	time	61	topic	176
degree	manner	12	topic	46	ARG1	164
compound	purpose	12	op2	46	compound	163
domain	theme	12	degree	44	name	160

Table 8: Error analysis on private test for Team UIT-VNS.

Gold	Pred	Count	Missing label	Count	Extra label	Count
theme	topic	70	domain	496	op1	414
agent	pivot	51	theme	187	agent	357
modality	tense	23	modality	185	op2	333
poss	ARG1	20	manner	123	domain	319
theme	patient	19	agent	101	compound	222
manner	degree	18	time	91	theme	221
theme	goal	16	quant	73	ARG2	186
domain	theme	14	topic	51	ARG1	165
agent	theme	13	purpose	43	name	139
theme	beneficiary	9	polarity	26	topic	126

Across all teams, **pivot-to-agent** and **theme-related** mispredictions were consistently among the most common errors, indicating challenges in correctly identifying core semantic roles. In addition, confusion between domain-specific labels such as *op1*, *op2*, and *compound* was prevalent, particularly for ViAMR and UIT_BlackCoffee, suggesting difficulties in capturing hierarchical or relational structures in Vietnamese semantic parsing.

Missing labels often involved critical semantic roles such as *domain*, *agent*, and *theme*, while extra labels were frequently inserted for *agent*, *op1*, and *domain*. This pattern implies that models tended to over-predict high-frequency roles while under-predicting less frequent but semantically important roles.

Overall, the error analysis reveals that semantic parsing in Vietnamese remains challenging, especially for distinguishing fine-grained semantic relations and complex role assignments. Future improvements may benefit from enhanced role disambiguation strategies, incorporation of syntactic cues, and structured reasoning mechanisms to better capture relational dependencies in Vietnamese sentences.

Table 9: Error analysis on private test for Team LangMind.

Gold	Pred	Count	Missing label	Count	Extra label	Count
pivot	agent	37	classifier	154	agent	526
compound	direction	16	quant	124	domain	395
compound	manner	14	domain	119	op1	367
pivot	domain	10	agent	109	theme	349
topic	theme	10	polarity	104	op2	333
manner	degree	10	compound	88	topic	191
agent	theme	9	mode	84	ARG2	184
compound	theme	9	pivot	84	ARG1	171
theme	compound	8	degree	84	compound	171
op1	op2	7	op1	60	name	161

6 Conclusion

The Vietnamese Semantic Parsing (viSemParse) Shared Task showcased the potential of semantic parsing techniques in capturing deep meaning representations for Vietnamese. Over the course of several months, the competition attracted strong participation from the research community, reflecting growing interest in semantic understanding for low-resource languages. The gold-standard dataset, consisting of approximately 2,500 annotated sentences, was designed to support model training and evaluation across training, public test, and private test phases.

Among all submissions, the top-performing system achieved a Smatch score of 58%, demonstrating the feasibility of applying modern transformer-based approaches to Vietnamese semantic parsing. While the results remain below those observed for high-resource languages, they provide valuable insights and establish a solid foundation for future work on semantic representation and meaning-based NLP for Vietnamese.

Acknowledgments

We would like to express our sincere thanks to the annotation team for their careful and dedicated work in building and verifying the Vietnamese semantic parsing dataset. We also thank all participating teams for their valuable efforts, insightful ideas, and enthusiasm throughout the shared task.

This work was made possible thanks to the

support of the VLSP community and collaborating institutions that provided computational and organizational resources.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio C. T. Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025. Phi-4-reasoning: A 14-b parameter model for complex reasoning. *arXiv preprint arXiv:2504.21318*. Phi-4-reasoning Technical Report.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. SPRING: A simple and effective method for abstract meaning representation parsing and generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL 2021)*, pages 355–366.

Johan Bos. 2013. The groningen meaning bank. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, page 2, Trento, Italy.

Deng Cai and Wai Lam. 2020. AMRBART: Pre-training sequence-to-sequence models for amr parsing with latent structural information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 4899–4906.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Jayeol Chun and Nianwen Xue. 2024. Uniform meaning representation parsing as a pipelined approach. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 40–52, Bangkok, Thailand. Association for Computational Linguistics.

Google DeepMind. 2025. Gemma 3 technical report. Open-weight multilingual and multimodal model with 128 k context window and support for over 140 languages. Available at https://blog.google/technology/developers/gemma-3/.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, page arXiv:2407.

Michael Wayne Goodman. 2020. Penman: An open-source library and tool for AMR graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 312–319, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR 2022)*.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Ha Linh and Huyen Nguyen. 2019. A case study on meaning representation for Vietnamese. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 148–153, Florence, Italy. Association for Computational Linguistics.

Ha My Linh, Nguyen Thi Minh Huyen, Vu Xuan Luong, Nguyen Thi Luong, Phan Thi Hue, and Le Van Cuong. 2020. VLSP 2020 shared task: Universal Dependency parsing for Vietnamese. In *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, pages 77–83, Hanoi, Vietnam. Association for Computational Lingustics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2022. Bartpho: Pre-trained sequence-to-sequence models for vietnamese. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*, page 601–608. Association for Computational Linguistics.

Minh Nguyen, Tuan Anh Nguyen, and Dat Quoc Nguyen. 2021. Vit5: Pretrained text-to-text transformer for vietnamese language generation. In *Proceedings of the 19th Conference of the Pacific Association for Computational Linguistics (PACLING 2021)*, pages 289–300. Springer.

Phuong-Thai Nguyen, Xuan-Luong Vu, Thi-Minh-Huyen Nguyen, Van-Hiep Nguyen, and Hong-Phuong Le. 2009. Building a large syntactically-annotated corpus of Vietnamese. In *Proceedings of*

the Third Linguistic Annotation Workshop (LAW III), pages 182–185, Suntec, Singapore. Association for Computational Linguistics.

Abend O. and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 228–238. Association for Computational Linguistics.

Volha Petukhova and Harry Bunt. 2008. LIRICS semantic role annotation: Design and evaluation of a set of data categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Michael Regan, Shira Wein, George Baker, and Emilio Monti. 2024. MASSIVE multilingual Abstract Meaning Representation: A dataset and baselines for hallucination detection. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 1–17, Mexico City, Mexico. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. Available at https://qwenlm.github.io/blog/qwen3/.