VNU Journal of Science: Comp. Science & Com. Eng., Vol. 41, No. 2 (2025) 86–99

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

VLSP 2025 challenge: Vietnamese Semantic Parsing

Ha My Linh*, Pham Thi Duc, Le Ngoc Toan, Nguyen Thi Minh Huyen

[1] *VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam*

Received 06th December 2025;
Revised 10th December 2025; Accepted 22nd December 2025

Abstract: In 2025, the Eleventh Workshop on Vietnamese Language and Speech Processing (VLSP 2025) introduced its first shared task on Vietnamese Semantic Parsing, known as viSemParse. This task aims to assess how effectively participating systems can represent the deep semantic structure of Vietnamese sentences. To support model development and evaluation, the organizers created high-quality, task-specific annotated datasets.

The viSemParse 2025 corpus comprises 2,500 Vietnamese sentences, carefully partitioned into training, public test, and private test splits to support fair and reproducible evaluation. The shared task was conducted on the AIHub platform, where teams were required to submit predictions on the public test set before receiving their final ranking based on the hidden private test set, ensuring robustness against overfitting.

The best-performing system in the viSemParse track achieved a Smatch score of 58%, a result that highlights not only the inherent complexity of semantic parsing in Vietnamese but also the substantial opportunities for methodological advances and future research in this area.

Keywords: Vietnamese semantic parsing, viSemParse, VLSP 2025

1. Introduction

Semantic parsing plays a central role in natural language understanding, as it seeks to map sentences into structured representations that explicitly capture their intended meaning. In recent years, a variety of semantic representation frameworks and annotated datasets have been proposed to model meaning at different levels, ranging from individual words to sentences and larger discourse units. These representations support robust language interpretation across diverse contexts while also addressing key challenges such as ambiguity and semantic underspecification.

Among the well-established semantic resources, PropBank [1] offers role-based predicate–argument annotations, whereas Abstract Meaning Representation (AMR) [2] provides a deeper, graph-based encoding of sentence-level meaning.

*Corresponding author.
E-mail address: linhhm@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6493

In addition, several other corpora and annotation frameworks have been developed, including the Groningen Meaning Bank (GMB) [3], Universal Conceptual Cognitive Annotation (UCCA) [4], and Uniform Meaning Representation (UMR),...

Within these formalisms, Abstract Meaning Representation (AMR) [2] has gained substantial attention because it captures predicate–argument relations, semantic roles, and concept-level links in a unified graph structure. Research in AMR parsing for English has advanced rapidly, driven by neural encoder–decoder models and the emergence of large language models (LLMs). Notable systems include SPRING [5], which builds on the T5 architecture [6], and AMRBART [7], which adapts BART with graph-centric pretraining strategies. These models have pushed the state of the art, producing highly accurate graph-based semantic representations.

For Vietnamese, a number of studies have investigated how AMR can be adapted for semantic parsing. Linh et al. [8] introduced adjustments to the AMR formalism to better reflect structural characteristics specific to Vietnamese, providing an initial foundation for AMR-style annotation in this language. More recently, Regan et al. [9] presented MASSIVE-AMR, a large-scale resource comprising more than 84,000 text-to-graph annotations-the most extensive and linguistically diverse AMR dataset to date. It includes AMR representations for 1,685 information-seeking utterances spanning over 50 languages, including Vietnamese. Constructing high-quality corpora with semantic role annotations and developing associated tools is essential for supporting low-resource languages like Vietnamese and for fostering progress within the Vietnamese NLP research community.

The VLSP 2025 Shared Task on Vietnamese Semantic Parsing (viSemParse) represents the first large-scale benchmark dedicated to AMR-style semantic parsing for Vietnamese. The task provides a manually curated dataset of 2,500 sentences in PENMAN format [10], divided into training, public test, and private test splits. Its main objective is to assess system capabilities in modeling semantic relations and preserving structural coherence while also promoting research on multilingual transfer and semantic graph generation for low-resource settings. Participants were tasked with building systems that generate AMR-like semantic graphs for Vietnamese input sentences and submitting their outputs for evaluation using the Smatch metric [11].

Participating teams explored a broad set of modeling strategies. Some systems fine-tuned sequence-to-sequence transformer models on the viSemParse dataset to learn direct mappings from sentences to linearized AMR graphs, while others adapted Large Language Models-such as Qwen3-14B [12], Gemma-3 [13], and Phi-4 [14]-through instruction tuning or LoRA [15] for more efficient training. Several submissions also incorporated dedicated pipelines with data normalization, variable recovery, and rule-based post-processing to improve structural accuracy. Together, these approaches demonstrate notable progress in Vietnamese semantic representation and provide a solid foundation for future low-resource semantic parsing research. This paper provides an overview of the viSemParse Shared Task, including the task design, dataset, participating systems, and evaluation methodology, and discusses its importance for advancing Vietnamese semantic parsing and developing language understanding applications in low-resource settings. As the first large-scale benchmark for AMR-style semantic parsing in Vietnamese, this initiative represents a key step toward deeper semantic modeling and cross-lingual transfer in Vietnamese NLP.

2. viSemParse Shared Task

This section provides a detailed description of the ViSemParse shared task and the process of

constructing the dataset.

2.1. Shared Task Description

The goal of developing a Vietnamese semantic parser is to enable accurate understanding and formal representation of Vietnamese sentences by analyzing their syntactic and semantic structures. The system extracts underlying meanings and converts them into structured forms such as AMR or logical representations. This shared task centers on building a Vietnamese benchmark with semantic annotations and evaluating the performance of semantic parsing models.

For example, figure 1 presents the semantic graph for the Vietnamese sentence "Anh nói rõ cho em nghe thử coi." The predicate *nói* (say) forms the root of the graph, with *Anh* as its agent (`:agent`) and *rõ* (clearly) as its manner (`:manner`). A purpose relation (`:purpose`) links this event to the subordinate action *em nghe thử coi* (to listen). The resulting structure abstracts away from surface syntax and captures the core meaning, emphasizing how actions, participants, and intentions relate to one another.

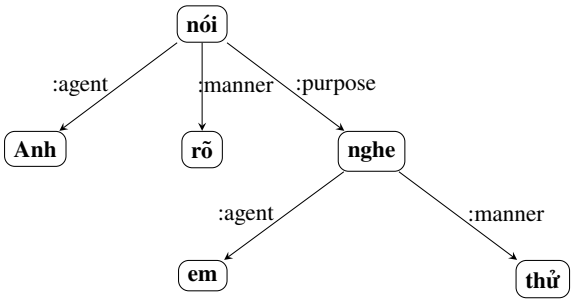

Figure 1. Semantic representation for the sentence "*Anh nói rõ cho em nghe thử coi.*"

So, a Vietnamese semantic parser functions in the following way:

- *Input*: A Vietnamese natural language sentence.
- *Output*: A semantic graph encoded in PENMAN format [10], with nodes denoting concepts and edges capturing their semantic relations.

The Smatch metric is widely used to evaluate the performance of semantic parsing systems. Smatch[1] [11] 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 semantic alignment between predicted and gold AMR graphs was evaluated using the Smatch F_1 metric, where higher scores indicate a stronger correspondence in meaning.

For example, we have two sentences in PENMAN format as in Figure 2.

In this case, the two AMR graphs are almost identical: they share the same set of concepts and match on 10 out of 11 triples. The only difference lies in one relation, where the gold graph uses `:manner` while the system output uses `:mod` for the same node.

Because there is just a single mismatched relation, so:

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

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

(a) Gold AMR representation

```
(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."

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

The shared task took place on the AIHub platform[2] and followed a two-stage format. During the public test phase, teams iteratively submitted outputs to track their progress on the leaderboard. The private test phase was then used to finalize system performance, and the overall rankings were computed based on these scores together with the submitted technical descriptions.

2.2. Data preparation

The Vietnamese Semantic Parsing dataset was developed through a structured two-phase workflow aimed at guaranteeing linguistic soundness and consistent annotations. This process focused on producing a high-quality corpus mapped to the AMR framework while carefully accommodating the distinctive syntactic and semantic properties of Vietnamese.

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

Table 1. Comparison of triples between the Gold and System AMR graphs

Type	Gold AMR Triple	System AMR Triple
Instance	(n, instance, nói)	(n, instance, nói)
Instance	(a, instance, Anh)	(a, instance, Anh)
Instance	(r, instance, rõ)	(r, instance, rõ)
Instance	(n1, instance, nghe)	(n1, instance, nghe)
Instance	(e, instance, em)	(e, instance, em)
Instance	(t, instance, thử)	(t, instance, thử)
Relation	(n, agent, a)	(n, agent, a)
Relation	(n, manner, r)	(n, manner, r)
Relation	(n, purpose, n1)	(n, purpose, n1)
Relation	(n1, pivot, e)	(n1, pivot, e)
Relation	**(n1, manner, t)**	**(n1, mod, t)**
Attribute	(root, TOP, n)	(root, TOP, n)

2.2.1. Vietnamese Semantic Labels

To build the Vietnamese semantic label set, we examined how meaning is conveyed differently in English versus Vietnamese. It became clear that introducing extra semantic labels was necessary to represent those unique Vietnamese nuances. The aim of our semantic representation model goes beyond answering the basic "who does what to whom" - it also seeks to encode where, when, why, and how. The core semantic roles in the Vietnamese model draw on both LIRICS [16] and English AMR [2], but the label set was extended further: it incorporates mechanisms for handling co-reference, tense–aspect, and additional categories to represent function words and modifiers - thereby addressing some of AMR's original limitations.

The Vietnamese semantic labels consist of 29 core roles, 88 non-core roles and 5 sentence-type labels. The main labels in the Vietnamese semantic representation include:

- *Predicates*: in Vietnamese, the predicate is a central component of the sentence structure, expressing an event, an action, a state, or a process. Predicates are most commonly realized by verbs, though in many cases nouns, adjectives, or modal verbs can also function as predicates depending

on the syntactic and semantic context. The predicate works together with other constituents - such as the subject, object, and adverbial modifiers-to form a complete and coherent proposition.

When defining predicates for the Vietnamese semantic representation, different types of predicates were carefully examined and separated into finer subcategories. For instance, modal verbs in Vietnamese such as "có thể" (can), "muốn" (want), "phải" (must), "khả năng" (likely), "nên" (should), . . . , are still treated as semantic predicate heads. This reflects the speaker's intention to describe degrees of necessity, obligation, permission, or possibility associated with an event.

A representative example of a Vietnamese predicate is illustrated below:

```
#::snt Ta có thể giúp đỡ cậu .
(p / possible-01
    :topic (g / giúp đỡ-02
        :agent (t / ta))
        :beneficiary (c/ cậu))
```

- *Core roles*: The core semantic roles in the Vietnamese semantic representation model are adapted and integrated from both LIRICS [16] and the English AMR framework [2]. The final inventory consists of 29 core role labels. These roles, along with their detailed mappings to LIRICS and AMR counterparts, are presented in Table **??**. They serve as the foundation for encoding fundamental participant relationships, including 29 roles such as *agent*, *patient*, *theme*, *beneficiary*, *goal*, *time*, and *location*, . . .
- *Non-core roles*: The Vietnamese semantic representation model includes 88 non-core semantic roles, covering a wide range of adjunct meanings such as comitative participants, beneficiaries, age, conditions, degree, destination, direction, instrument, location, and manner. These labels were selected to capture semantic phenomena commonly found in Vietnamese texts. Several language-specific cases are treated as follows:

  – Classifiers: Vietnamese classifiers (e.g., "cái", "chiếc", "quyển") precede nouns to mark categorization. When a classifier refers back to a previously mentioned noun, it still conveys recoverable semantic information. These are encoded using *:classifier*.

  For example:

  ```
  (s/sách
      :classifier (q/quyển))
  ```

  – Tense and aspect markers: While English AMR does not encode tense/aspect directly, Vietnamese uses particles such as "đã", "đang", "sẽ" to indicate temporal interpretation. Because their meaning depends heavily on context, the model assigns them the label *:tense* even when their temporal reference shifts. Example:

  ```
  (l/làm
      :agent (t/tôi)
      :tense (s/sẽ))
  ```

  – Compound relation: Multi-morphemic or multi-word expressions that form a new meaning (e.g., "ăn uống", "đi bộ") use the *:compound* relation.

  ```
  (n/nhảy
      :compound (m/múa))
  ```

  – Temporal expressions: Various adverbials of time are mapped to the appropriate *:time* labels (18 labels), including always, sometimes, now, before, and after.

– Named Entities: Named entities in the Vietnamese semantic model follow English AMR conventions. The *:wiki* attribute links an entity to its Wikipedia entry, and *:name* structures the surface name.

```
(p / person
    :wiki "Hồ_Chí_Minh"
    :name (n / name
        :op1 "Hồ"
        :op2 "Chí"
        :op3 "Minh"))
```

– Co-reference: The Vietnamese semantic representation supports document-level co-reference, similar to the multi-sentence co-reference annotations introduced in some English AMR documents after 2018. Sentences and tokens are assigned incremental IDs, and referential links are defined across an entire paragraph, supporting applications such as summarization, semantic analysis, and question answering.

– Sentence Types: Additional labels capture Vietnamese sentence types, including imperatives, exclamatives, interrogatives, compound sentences, and unknown-question forms. For multi-clause sentences without conjunctions, the label *:multi-sentence* is used along with attributes such as *:snt1*, *:snt2*, ...

Full definitions of the labels and detailed annotation guidelines are provided in the annotation manual[3].

2.2.2. Data Annotation

In the second phase, manual semantic annotation was carried out over a six-month period by a team of five linguistics experts. Each batch of data was independently annotated by two annotators to ensure reliability, after which the annotations were systematically cross-checked and any disagreements were resolved through discussion. This multi-stage annotation and review process helped ensure semantic correctness as well as a high level of inter-annotator agreement.

To facilitate the annotation workflow, a dedicated web-based platform was developed. The tool supports sentence visualization, interactive manipulation of AMR graphs, and automatic validation of relation types and labels references. These features enabled annotators to work in a consistent and efficient manner throughout the annotation process.

Quality assurance mechanisms were incorporated throughout the process. Inter-annotator agreement (IAA) was evaluated using Smatch [11] on a representative portion of the data, and disagreements were resolved through team consensus. In addition, automated scripts detected issues such as variable conflicts, missing relations, and structural errors. Through these combined procedures, the final corpus achieved high semantic quality and stable annotation consistency, with detailed agreement statistics reported in Table 2.

Table 2. 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**

2.2.3. Vietnamese Semantic Corpus

The corpus was constructed using several reliable linguistic resources, including the VietTreebank (VTB) [17], The Little Prince, and

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

the Vietnamese Dependency Parsing dataset [18]. Together, these sources offer extensive syntactic and semantic variation, allowing the resulting dataset to capture a broad range of real-world Vietnamese language patterns. Detailed corpus statistics are summarized in Table 3, which reports the distribution of instances across the training, public test, and private test subsets.

Table 3. Statistics of the Vietnamese Semantic Parsing dataset

Dataset	# 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

Table 4 reports the frequency of several representative labels in the corpus. Core semantic roles such as *:mod*, *:agent*, and *:theme* appear most frequently, reflecting their importance in encoding fundamental predicate-argument relations. Non-core roles - including *:degree*, *:manner*, *:time*, and various *:op* arguments - also occur regularly, indicating that the corpus effectively captures contextual information related to manner, temporality, and intensity. Overall, the distribution of labels shows balanced coverage of both core and non-core roles, forming a robust basis for training Vietnamese semantic parsing models.

Table 4. Most frequent semantic labels in the training dataset

Label	Fre.	Label	Fre.
:mod	1,236	:domain	392
:agent	1,142	:op1	381
:theme	776	:op2	362
:pivot	541	:polarity	299
:compound	534	:time	290
:topic	487	:name	288
:classifier	448	:degree	425
:quant	416	:manner	394

3. Vietnamese Semantic Parsing Methods

In the 2025 edition of viSemParse, a total of 12 teams participated, generating 338 system runs across the public and private evaluation stages. This section first describes the baseline system and then reviews the techniques and modeling strategies reported in the teams' system description papers, providing a consolidated view of how participants approached the Vietnamese semantic parsing task.

3.1. Baseline System

For the baseline system, we formulate Vietnamese AMR parsing as a conditional sequence-to-sequence generation task. We choose the Qwen3-1.7B causal language model as our backbone, leveraging its strong generative capabilities. To allow efficient adaptation under limited computational resources, we fine-tune the model using QLoRA, which enables low-rank parameter updates while keeping the majority of the model frozen. The model is loaded in 4-bit quantization to reduce memory footprint and speed up training without significantly affecting performance. LoRA adapters are carefully inserted into all main attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) as well as the feed-forward layers (`gate_proj`, `up_proj`, `down_proj`), which we found crucial for capturing long-range dependencies and complex structural patterns in Vietnamese AMR graphs.

Training is conducted for three epochs with an effective batch size of 512. We use AdamW optimization with a learning rate of 1×10^{-5} and a cosine learning rate scheduler. The maximum sequence length is set to 2048 tokens to accommodate long Vietnamese sentences and deeply nested AMR structures. These hyperparameters were selected empirically based on stability during training and the ability to handle typical sentence lengths in our dataset.

To guide the model towards producing linguistically well-formed AMRs, we employ

a fixed instruction-style prompt. The prompts are carefully designed to provide clear and unambiguous input-output specifications, guiding the model to produce well-structured AMRs. They include detailed definitions of semantic labels specific to Vietnamese, highlight important linguistic features such as multiword expressions and role distinctions, and point out particular cases or exceptions that require special attention to ensure accurate graph construction. Additionally, structural constraints are embedded in the prompt to maintain balanced parentheses and prevent the inclusion of natural-language explanations in the generated output. During inference, AMRs are generated with a low temperature (0.2) and a 256-token limit to reduce the risk of structural drift. Generated outputs undergo lightweight post-processing to remove prompt artifacts and ensure basic Penman well-formedness.

To enhance transparency and reproducibility, we fully document all aspects of the training setup, including hyperparameters, prompt design, LoRA insertion points, quantization details, and inference configuration. This detailed description allows other researchers to reproduce our baseline and provides a foundation for systematic ablation studies. For example, future analyses could measure the impact of different prompt designs, the contribution of each LoRA insertion location, or the effects of quantization and sequence length limits on structural accuracy.

Overall, this configuration provides a computationally efficient, stable, and transparent baseline that captures the core requirements of Vietnamese AMR graph generation, while leaving room for systematic investigation into the contributions of individual design choices.

3.2. Participant Approaches

Table 5 summarizes the techniques employed by the top-performing teams, with the subsequent subsections providing a more detailed explanation of each approach. The participating systems varied widely in design, drawing on both large pre-trained models and specialized fine-tuning methods. Several teams adapted recent transformer-based LLMs - such as Qwen3, Gemma-3, LLaMA-3.1, and Phi-4 - using LoRA or full supervised tuning to tailor them to the semantic parsing task in Vietnamese. Other teams experimented with alternative solutions, including multi-agent pipelines and Vietnamese-oriented architectures like BARTpho or ViT5. Collectively, these approaches illustrate a thoughtful blend of cutting-edge LLM adaptation techniques and linguistically motivated modeling, signaling continued progress in Vietnamese NLP research.

Table 5. 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

Next, we present a detailed account of the approaches, architectures, and techniques used by each participating team, highlighting the diversity of strategies applied to Vietnamese semantic parsing.

Team UIT_BlackCoffee: The proposed approach implements a three-stage pipeline for Vietnamese semantic parsing leveraging large language models (LLMs), including Qwen3-14B [12], Gemma-3 [13], Phi-4 [14], and LLaMA-3.1 [19]. These models are fine-tuned using Supervised Fine-Tuning (SFT) [20] with LoRA [15] to reduce computational overhead.

In preprocessing, non-semantic elements

such as variables and wiki tags are removed, and AMR graphs are linearized into PENMAN-style sequences to facilitate model training. Following fine-tuning, a post-processing step restores variables and assigns unique identifiers to concept nodes, ensuring structural validity and producing well-formed AMR graphs.

By employing a 4-bit quantized Qwen3-14B model with AdamW optimization and early stopping, this method effectively adapts multilingual LLMs for Vietnamese semantic parsing, achieving improved graph accuracy and maintaining syntactic consistency across outputs.

Team ViAMR: The approach utilizes a three-stage pipeline for Vietnamese AMR parsing, integrating preprocessing, supervised fine-tuning, and a constraint-aware inference process. During preprocessing, PENMAN-formatted graphs are normalized into single-line sequences, missing brackets are corrected, and multiword nodes are joined with underscores to preserve syntactic correctness.

For model training, a compact decoder-only LLM (Qwen3-1.7B [12]) is fine-tuned using Supervised Fine-Tuning (SFT) [20] to map Vietnamese sentences to linearized AMR graphs in an instruction-following setup. Optimization leverages AdamW with linear learning rate decay, gradient accumulation, and distributed training to ensure efficient convergence.

At inference, the system generates PENMAN strings and performs a series of repairs at both string and graph levels, including role spacing adjustments, bracket balancing, variable deduplication, and canonicalization through PENMAN round-trip parsing. This workflow produces well-formed, structurally consistent AMR graphs with stable evaluation performance while remaining efficient on limited computational resources.

Team UIT-VNS: The ViSemCrew framework approaches Vietnamese semantic parsing using a multi-agent workflow, breaking the overall task into specialized subtasks handled by different agents. It consists of five coordinated components: the Linguistic Analysis Agent, responsible for morphological and syntactic processing; the Concept Extraction Agent, which identifies semantic concepts; the Graph Construction Agent, in charge of establishing relations and selecting the root predicate; the Validation Agent, which ensures the correctness of the generated graph; and the Repair Agent, which detects errors and performs regeneration when necessary.

These agents operate in a sequential manner, incorporating iterative validation and fallback mechanisms to improve reliability. The system also incorporates Vietnamese-specific adaptations, including passive voice handling, flexible word order management, and recovery of implicit elements. In addition, a role reference database supports the agents, enhancing both the accuracy and consistency of semantic parsing.

Team LangMind: The proposed approach leverages an encoder–decoder transformer framework to generate AMR-style semantic graphs from Vietnamese sentences. It fine-tunes two pretrained Vietnamese language models, BARTpho [21] and ViT5 [22], on the official VLSP 2025 dataset. The preprocessing pipeline involves sentence tokenization and normalization, linearization of graphs via depth-first traversal, and filtering out inconsistent AMRs to ensure high-quality input.

The system investigates tokenization at both word and syllable levels, treating punctuation as separate tokens to maintain structural fidelity. During training, the models are optimized using AdamW with a learning rate of 5×10^{-5} and a batch size of 8, with early stopping applied. At inference, beam search with a beam size of 4 is employed to generate candidate sequences. Post-processing includes normalizing variable names, correcting bracket mismatches, and reconstructing AMR graphs from the linearized sequences.

Experimental results show that BARTpho

with word-level tokenization achieves the best performance, indicating that fine-grained word segmentation significantly enhances graph connectivity and semantic accuracy in Vietnamese AMR parsing.

4. Results and Discussion

This section summarizes the outcomes of the participating systems, examining their performance in depth, identifying frequent types of errors, and discussing key lessons for enhancing future models.

4.1. Results

Table 6 summarizes the performance of the participating teams and the baseline on the public test set, which included 164 submissions. UIT_BlackCoffee achieved the highest Smatch score of 0.55, with a precision of 0.53 and a recall of 0.57, reflecting a well-balanced ability to capture both frequent and complex AMR structures. UIT-VNS followed with a Smatch of 0.42, showing competitive performance but still lagging behind the leader.

ViAMR reached a moderate Smatch of 0.38, while LangMind scored 0.33, with high precision (0.46) but low recall (0.26), indicating under-prediction of several labels. The baseline system obtained a Smatch of 0.44, situating it between top-performing and mid-tier teams. Overall, these results highlight differences in how systems balance precision and recall, and underscore the challenges in parsing complex Vietnamese AMR structures.

Table 6. 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
-	Baseline	0.42	0.46	0.44

During the private test phase, 174 submissions were evaluated, with some shifts in team rankings compared to the public test. Table 7 summarizes the results, including the baseline for reference. UIT_BlackCoffee led the evaluation with a Smatch score of 0.58, achieving strong precision (0.52) and recall (0.64), which indicates excellent generalization to unseen data.

ViAMR and UIT-VNS maintained relatively stable performance, scoring 0.46 and 0.42, respectively, showing consistent results but still below the leading team. LangMind improved slightly to 0.37, reflecting moderate gains in recall (0.42), yet it remained the lowest-ranking system.

Table 7. 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
-	Baseline	0.46	0.52	0.48

Overall, these results highlight UIT_BlackCoffee's robustness across datasets, while UIT-VNS performs well on familiar structures but demonstrates less adaptability. Both LangMind and ViAMR would benefit from further improvements in precision and recall to match the performance of the top teams.

4.2. Errors Analysis

A comparative examination of the private test set reveals several consistent patterns in the types of semantic parsing errors produced by the four leading teams - UIT_BlackCoffee, ViAMR, UIT-VNS, and LangMind. Instead of describing each table sequentially, this section synthesizes the findings into broader error categories, drawing attention to cross-team tendencies and system-specific weaknesses.

4.2.1. Substitution Errors

Table 8 shows the top substitution errors across four teams, highlighting cases where the model predicts an incorrect label.

Table 8. Top substitution errors across four teams

Team	Gold → System	Count
UIT_BlackCoffee	pivot → agent	52
UIT_BlackCoffee	topic → theme	25
UIT_BlackCoffee	compound → direction	22
ViAMR	pivot → agent	39
ViAMR	compound → manner	25
ViAMR	compound → direction	22
UIT-VNS	theme → topic	70
UIT-VNS	agent → pivot	51
UIT-VNS	modality → tense	23
LangMind	pivot → agent	37
LangMind	compound → direction	16
LangMind	compound → manner	14

The most frequent error pattern concerns the *pivot* label, which is consistently mispredicted as *agent* in UIT_BlackCoffee, ViAMR, and LangMind, with misclassification counts ranging from 37 to 52. This systematic confusion indicates that the model struggles to distinguish between pivot and agent roles, particularly in constructions where pivot arguments co-occur with agents or occupy structurally central positions in the clause. In Vietnamese, pivot roles are often realized without explicit morphological marking and may share similar syntactic distributions with agents, making their distinction heavily dependent on broader semantic and discourse context. As a result, models that primarily rely on surface-level syntactic cues or positional information are prone to conflating these two roles, especially in sentences where contextual signals are subtle or underspecified.

Another common source of errors is the *compound* label, which is frequently mispredicted as direction or manner. This reflects the challenge of modeling composite semantic roles, where a single label can encompass multiple aspects of meaning such as direction, manner, or thematic relations. Similarly, substitutions between theme and topic are prevalent, particularly in UIT_BlackCoffee and UIT-VNS, indicating that the model struggles to differentiate these closely related thematic roles.

Some team-specific substitutions are also notable, such as *agent* → *pivot* and *modality* → *tense* in UIT-VNS, reflecting ambiguity in semantic or grammatical cues. Overall, substitution errors are concentrated on central, composite, or semantically complex labels, highlighting the need for better context-aware modeling to reduce confusion among these roles.

4.2.2. Missing Label Errors

Missing label errors arise when the model fails to predict a label that should be present, as shown in Table 9. Among the four teams, the labels most commonly omitted are *domain*, *agent*, and *theme*. Notably, UIT-VNS exhibits a particularly high omission of domain with 496 instances, suggesting that the model has difficulty identifying central semantic roles in some sentences.

These errors can often be attributed to ambiguity in the input data or to limited coverage in the training set. In particular, labels such as *theme* and *agent* are sometimes only implicitly realized in the text, which makes them difficult for the model to identify consistently. In addition, the relatively small number of missing-label cases for certain roles, such as *classifier* in LangMind, suggests that team-specific annotation conventions and data characteristics also play a role in shaping model performance.

Overall, missing label errors highlight that the models tend to under-predict labels with central or composite semantic roles, which can significantly affect downstream interpretation and reasoning tasks in AMR parsing.

Table 9. Top missing label errors across four teams

Team	Missing Label	Count
UIT_BlackCoffee	domain	92
UIT_BlackCoffee	agent	57
UIT_BlackCoffee	compound	49
ViAMR	domain	113
ViAMR	agent	88
ViAMR	compound	74
UIT-VNS	domain	496
UIT-VNS	theme	187
UIT-VNS	modality	185
LangMind	classifier	154
LangMind	quant	124
LangMind	domain	119

Table 10. Top extra label errors across four teams

Team	Extra Label	Count
UIT_BlackCoffee	agent	474
UIT_BlackCoffee	op1	306
UIT_BlackCoffee	domain	269
ViAMR	agent	481
ViAMR	op1	363
ViAMR	domain	352
UIT-VNS	op1	414
UIT-VNS	agent	357
UIT-VNS	op2	333
LangMind	agent	526
LangMind	domain	395
LangMind	op1	367

4.2.3. Extra Label Errors

Table 10 presents the top extra label errors across four teams, showing cases where the model predicts labels that are not present in the gold standard.

The agent label is the most frequently over-predicted, with counts ranging from 357 in UIT-VNS to 526 in LangMind. This suggests that models often assign the agent role too broadly, even in cases where agency is weak, implicit, or not semantically justified. In Vietnamese, agent roles are typically inferred from word order or contextual cues rather than explicit morphological markers. As a result, when models encounter prominent verbs or animate entities, they tend to default to labeling them as agent, leading to frequent over-prediction.

Other labels that are frequently over-predicted include *op1* and *domain*. Specifically, UIT_BlackCoffee predicts *op1* 306 times and *domain* 269 times when they are not present in the gold annotations, while ViAMR shows an even stronger tendency, adding *op1* 363 times and *domain* 352 times. This pattern suggests that the models rely heavily on label frequency or prominent lexical cues in the input, which leads to many false positives for these high-frequency semantic roles.

5. Conclusion

The Vietnamese Semantic Parsing (viSemParse) Shared Task demonstrates that recent semantic parsing methods are increasingly capable of modeling deep and structured meaning representations for Vietnamese, despite the language's limited annotated resources. The competition was conducted over several months and attracted wide participation from both academic and industrial research teams, reflecting a growing interest in semantic understanding for low-resource languages. The task was based on a gold-standard dataset of around 2,500 manually annotated sentences, divided into training, public test, and private test sets. This dataset covers a broad range of syntactic and semantic phenomena in Vietnamese, providing a realistic and challenging benchmark for evaluating system performance.

Among all submissions, the best-performing system achieved a Smatch score of 58%, showing that approaches leveraging large language models can capture Vietnamese semantic structures reasonably well despite the limited amount of annotated data. Performance varied across the remaining teams, reflecting the difficulty of modeling implicit information and language-specific phenomena. Although the overall scores are still lower than those typically reported for

high-resource languages such as English, the results offer useful insights into the current capabilities and limitations of existing methods.

Overall, the viSemParse establishes a benchmark for Vietnamese semantic parsing and offers a solid foundation for further research. The insights gained from system performance, error analysis, and dataset challenges can guide the development of more robust, context-aware, and linguistically informed models, ultimately advancing meaning-based NLP applications 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

[1] P. Kingsbury, M. Palmer, From TreeBank to PropBank, in: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain, 2002.

[2] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, N. Schneider, Abstract Meaning Representation for Sembanking, in: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 178–186.

[3] J. Bos, The Groningen Meaning Bank, in: Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora, Trento, Italy, 2013, p. 2.
URL https://www.aclweb.org/anthology/W13-3802

[4] O. Abend, A. Rappoport, Universal Conceptual Cognitive Annotation (UCCA), in: Proceedings of ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1, Association for Computational Linguistics, 2013, pp. 228–238.

[5] M. Bevilacqua, R. Blloshmi, R. Navigli, 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), 2021, pp. 355–366.
URL https://aclanthology.org/2021.eacl-main.28

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research, Vol. 21, No. 140, 2020, pp. 1–67.
URL https://jmlr.org/papers/v21/20-074.html

[7] D. Cai, W. Lam, 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), 2020, pp. 4899–4906.
URL https://aclanthology.org/2020.emnlp-main.396

[8] H. Linh, H. Nguyen, A Case Study on Meaning Representation for Vietnamese, in: Proceedings of the First International Workshop on Designing Meaning Representations, Association for Computational Linguistics, Florence, Italy, 2019, pp. 148–153.

[9] M. Regan, S. Wein, G. Baker, E. Monti, MASSIVE Multilingual Abstract Meaning Representation: A Dataset and Baselines for Hallucination Detection, in: D. Bollegala, V. Shwartz (Eds.), Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024), 2024.

[10] M. W. Goodman, Penman: An Open-Source Library and Tool for AMR Graphs, in: A. Celikyilmaz, T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 312–319. doi:10.18653/v1/2020.acl-demos.35.
URL https://aclanthology.org/2020.acl-demos.35/

[11] S. Cai, K. Knight, 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), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 748–752.
URL https://www.aclweb.org/anthology/P13-2131

[12] Q. Team, Qwen3 Technical Report, available at https://qwenlm.github.io/blog/qwen3/ (2025).

[13] G. DeepMind, 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/ (2025).

[14] M. Abdin, S. Agarwal, A. Awadallah, V. Balachandran, H. Behl, L. Chen, G. de Rosa, S. Gunasekar, M. Javaheripi, N. Joshi, P. Kauffmann, Y. Lara, C. C. T. Mendes, A. Mitra, B. Nushi, D. Papailiopoulos, O. Saarikivi, S. Shah, V. Shrivastava, V. Vineet, Y. Wu, S. Yousefi, G. Zheng, Phi-4-Reasoning: A 14-B Parameter Model for Complex Reasoning, arXiv preprint arXiv:2504.21318Phi-4-Reasoning Technical Report (2025).

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, in: Proceedings of the International Conference on Learning Representations (ICLR 2022), 2022.
URL https://openreview.net/forum?id=nZeVKeeFYf9

[16] V. Petukhova, H. Bunt, LIRICS Semantic Role Annotation: Design and Evaluation of a Set of Data Categories, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (Eds.), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008.
URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/17_paper.pdf

[17] P.-T. Nguyen, X.-L. Vu, T.-M.-H. Nguyen, V.-H. Nguyen, H.-P. Le, Building a Large Syntactically-Annotated Corpus of Vietnamese, in: Proceedings of the Third Linguistic Annotation Workshop (LAW III), Association for Computational Linguistics, Suntec, Singapore, 2009, pp. 182–185.

[18] H. M. Linh, N. T. M. Huyen, V. X. Luong, N. T. Luong, P. T. Hue, L. V. Cuong, VLSP 2020 Shared Task: Universal Dependency Parsing for Vietnamese, in: Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing, Association for Computational Lingustics, Hanoi, Vietnam, 2020, pp. 77–83.

[19] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The Llama 3 Herd of Models, arXiv e-prints 2024, pp. arXiv:2407.

[20] G. Dong, H. Yuan, K. Lu, C. Li, M. Xue, D. Liu, W. Wang, Z. Yuan, C. Zhou, J. Zhou, How Abilities in Large Language Models are affected by Supervised Fine-Tuning Data Composition, arXiv preprint arXiv:2310.05492 (2023).
URL https://arxiv.org/abs/2310.05492

[21] D. Q. Nguyen, A. T. Nguyen, BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022, Association for Computational Linguistics, 2022, pp. 601–608.
URL https://aclanthology.org/2022.findings-acl.53

[22] M. Nguyen, T. A. Nguyen, D. Q. Nguyen, 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), Springer, 2021, pp. 289–300. doi:10.1007/978-981-16-8964-7_23.