**ANALYSEUR SÉMANTIQUE POUR LA LANGUE NATURELLE**

**SEMANTIC PARSER FOR NATURAL LANGUAGE**

*Etablissement* **Université de Lorraine**

*École doctorale* **IAEM - INFORMATIQUE - AUTOMATIQUE - ELECTRONIQUE - ELECTROTECHNIQUE - MATHEMATIQUES**

*Spécialité* **Informatique**

*Unité de recherche* **LORIA - Laboratoire Lorrain de Recherche en Informatique et ses Applications**

*Encadrement de la thèse* **Maxime AMBLARD**

**Financement** du 01-10-2025 au 30-09-2028

*Début de la thèse le* **1 octobre 2025**

*Date limite de candidature (à 23h59)* **24 avril 2025**

## Mots clés - Keywords

parser, sémantique, logique, graphes

parser, semantics, logic, graphs

## Description de la problématique de recherche - Project description

Le traitement du langage naturel a fait de grands progrès ces dernières années en produisant des modèles capables de résoudre des problèmes complexes. Si le traitement du langage naturel semble avoir atteint un bon niveau de compréhension, la question de la sémantique reste ouverte. Par exemple, le développement de solutions orientées vers les tâches nécessite l'extraction de la structure sous-jacente d'un énoncé afin d'en extraire les intentions. Cependant, la production de la structure prédicative d'un énoncé est au-delà de la capacité des LLM.

Deux grandes écoles de pensée ont émergé en sémantique des langues, l'une basée sur des propriétés logiques (Kamp et Reyle, 1993 ; Montague, 1970) et l'autre sur des représentations graphiques (Banarescu et al., 2013 ; Abend et Rappoport, 2013 ; White et al., 2016 ; Van Gysel et al., 2021). Alors que les premières sont plus adaptées à la construction d'analyseurs syntaxiques, les secondes permettent de considérer une quantité plus raisonnable de données certifiées, notamment si l'on souhaite utiliser des stratégies d'apprentissage automatique (Cheng 2017, Li 2018). Récemment, plusieurs initiatives ont été lancées pour désigner des formalismes de représentation sémantique dans lesquels les structures de graphe utilisées conservent de bonnes propriétés logiques (Michalon, 2016), dans le but de produire des corpus sémantiquement annotés, dont l'une des fonctions est d'entraîner des analyseurs syntaxiques. Au stade actuel de développement, la qualité obtenue est encore en deçà d'un niveau permettant son utilisation large.

Dans ce sujet, nous nous concentrons sur un type spécifique de représentation sémantique, le cadre YARN (Pavlova 2024) (laYered meAning RepresentatioN), avec une structure argumentale de prédicat (structure PA) basée sur la représentation abstraite du sens (AMR) (Banarescu et al., 2013) et une approche en couches pour encoder d'autres phénomènes sémantiques. Un intérêt majeur de cette représentation est qu'elle est basée sur une représentation simplifiée de la structure argumentale, qui reste plus accessible en termes de développement d'outils, tout en permettant de considérer une grande complexité de phénomènes sémantiques classiques, tels que la négation, la modalité, la temporalité et la quantification, et la manière dont ils peuvent interagir entre eux. Considérer les interactions entre différents phénomènes est un défi que les formalismes existants n'abordent pas explicitement.

Dans la tâche d'analyse sémantique, le système doit prédire la structure formelle qui représente les liens sémantiques. Les stratégies d'apprentissage automatique ont des difficultés à se généraliser à des structures qui ne sont pas présentes dans les grands ensembles de données utilisés pour l'entraînement. Ce phénomène bien connu a conduit à envisager de combiner plusieurs approches qui intègrent des représentations de structures sémantiques dans les architectures utilisées, notamment dans les codeurs et décodeurs (Petit 2023, Petit et Corro 2024).

L'objectif du stage est de construire un analyseur sémantique pour le formalisme YARN et de l'évaluer sur des données réelles comme Geoquery, graphQuestions, Spades, etc.

NLP has made great strides in recent years in producing models capable of solving complex problems. While natural language processing seems to have reached a good level of understanding, the question of semantics is still open. For example, the development of task-oriented solutions requires the extraction of the underlying structure of an utterance in order to extract its intentions. However,

producing the predicative structure of an utterance is beyond the capacity of LLMs.

Two main schools of thought have emerged in language semantics, one based on logical properties (Kamp and Reyle, 1993; Montague, 1970) and the other on graphical representations (Banarescu et al., 2013; Abend and Rappoport, 2013; White et al., 2016; Van Gysel et al., 2021). While the former are more suitable for the construction of parsers, the latter allow to consider a more reasonable amount of certified data, especially if we want to use machine learning strategies (Cheng 2017, Li 2018). Recently, several initiatives have been launched to designate semantic representation formalisms in which the graph structures used retain good logical properties (Michalon, 2016), with the aim of producing semantically annotated corpora, one of the functions of which is to train parsers. At the current stage of development, the quality obtained is still below a level that allows it to be used.

In this topic, we focus on a specific type of semantic representation, YARN (Pavlova 2024) framework (laYered meAning RepresentatioN), with a predicate argument structure (PA structure) based on the Abstract Meaning Representation (AMR) (Banarescu et al., 2013) and a layered approach to encode other semantic phenomena. A major interest of this representation is that it is based on a simplified representation of the argument structure, which remains more accessible in terms of tool development, while allowing us to consider a great complexity of classical semantic phenomena, such as negation, modality, temporality and quantification, and how they can interact with each other. Considering the interactions between different phenomena is a challenge that existing formalisms do not explicitly address.

In the semantic parsing task, the system must predict the formal structure that represents the semantic links. Machine learning strategies have difficulty generalising to structures that are not present in the large datasets used for training. This well-recognised phenomenon has led us to consider combining several approaches that integrate representations of semantic structures into the architectures used, in particular in encoders and decoders (Petit 2023, Petit and Corro 2024).
The aim of the internship is to build a semantic parser for the YARN formalism and to evaluate it on real data as Geoquery, graphQuestions, Spades, etc.

## Thématique / Domaine / Contexte

Traitement Automatique des Langue

Informatique

L'équipe-projet INRIA Sémagramme fait partie du centre Inria de l'université de Lorraine ainsi que du département Traitement automatique des langues et des connaissances du laboratoire LORIA de l'Université de Lorraine.
https://team.inria.fr/semagramme/fr/

## Objectifs

Développement d'un analyseur sémantique de données en langue naturelle fondé sur le formalisme YARN.

## Références bibliographiques

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 77–89, Vancouver, Canada. Association for Computational Linguistics.

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.

Jianpeng Cheng, Siva Reddy, Vijay Saraswat, and Mirella Lapata. Learning structured natural language representations for semantic parsing. arXiv preprint arXiv:1704.08387, 2017.
Hans Kamp and Uwe Reyle. 1993. From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory. Dordrecht. Kluwer.

Li Dong and Mirella Lapata. Coarse-to-fine decoding for neural semantic parsing. arXiv preprint arXiv:1805.04793, 2018.

Olivier Michalon, Corentin Ribeyre, Marie Candito, Alexis Nasr. Deeper syntax for better semantic parsing. Coling 2016 - 26th International Conference on Computational Linguistics, Dec 2016, Osaka, Japan. hal- 01391678

Richard Montague. 1970. English as a formal language. Logic and philosophy for linguists.
Siyana Pavlova, Maxime Amblard, Bruno Guillaume. YARN is All You Knit: Encoding Multiple Semantic Phenomena with Layers. The Fifth International Workshop in Designing Meaning Representation, May 2024, Turin, Italy. hal-04551796

Alban Petit and Caio Corro. 2023. On Graph-based Reentrancy-free Semantic Parsing. Transactions of the Association for Computational Linguistics, 11:703–722.
Alban Petit, Structured prediction methods for semantic parsing, Thèse de doctorat dirigée par
Yvon, François et Corro, Caio Informatique université Paris-Saclay 2024

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, ChuRen Huang, et al. 2021. Designing a uniform meaning representation for natural language processing. KI-Künstliche Intelligenz, 35(3-4):343–360.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1713–1723.

## Précisions sur l'encadrement - Details on the thesis supervision

Suivi hebdomadaire du travail par l'encadrant
Cours de formation complémentaires obligatoire organisés par l'Ecole Doctorale
Présentation des travaux en séminaires et conférences

## Conditions scientifiques matérielles et financières du projet de recherche

Participation au concours des contrats doctoraux. Pas de condition de sécurité spécifique.
Participation in the doctoral contract competition. No specific safety requirements.

## Objectifs de valorisation des travaux de recherche du doctorant : diffusion, publication et confidentialité, droit à la propriété intellectuelle,...

Publication dans des conférences et revues internationales du domain.
Publication in international conferences and journals of the field.

## Profil et compétences recherchées - Profile and skills required

Master en NLP, en informatique ou dans un domaine connexe.
Maîtrise des langages de programmation (Python) et des bonnes pratiques de codage
Compétences en conception d'algorithmes
Expérience en apprentissage profond
Maîtrise de la logique, des graphes et des représenations sémantiques des énoncés en langue naturelle
Capacité à travailler de manière autonome et à travailler en équipe
Excellentes compétences en anglais, à l'oral et à l'écrit
Master's degree in NLP , Computer Science or a related master program
Proficiency in programming languages (Python) and good coding practices
Skills in algorithm design
Experience in deep learning
Proficiency in logic, graphs and semantic representations of natural language utterances
Ability to work independently and also to work in a team
Excellent oral and written English skills

Dernière mise à jour le 12 mars 2025