

Master’s Thesis Internship — IRISA (Rennes)

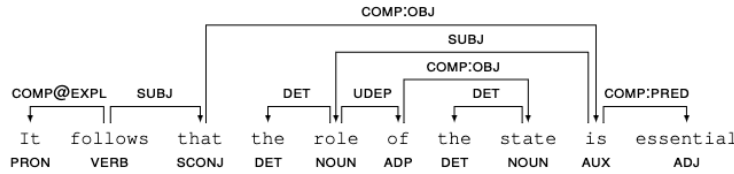
Computational Methods for Research in Formal Linguistics

Academic year 2026-2027

- Contract type: end-of-studies internship (4-6 months)
- Laboratory: IRISA, Rennes
- Research teams: Archimédia and Lacodam
- Supervisors:
Caio Corro, Peggy Cellier and Alexandre Termier
- Possibility to continue with a funded PhD



While natural language processing has long relied on a close collaboration between computer science and linguistics, in particular through the integration of linguistic knowledge to improve models, recent advances in the field have been achieved without any recourse to linguistics. The Autogramm AI project proposes to rethink this interdisciplinarity by aiming to advance research in formal linguistics using the tools of computer science, whether formal methods or machine learning methods, in the spirit of the “*AI for science*” approach that has become very popular in fields such as physics or climate science. We are particularly interested in the induction of descriptive grammars from annotated linguistic data [Herrera et al., 2024].



Example of a syntactically annotated sentence. Data following this annotation format are available for many languages through the Universal Dependencies project [Nivre et al., 2020, Gerdes et al., 2018].

Definition 1 (Syntactic grammar rule). Within all dependencies that match a pattern S (the scope), a grammar rule with frequency α identifies a predictor P that triggers linguistic phenomena Q in $\alpha\%$ of cases. We denote such a rule by:

$$S \Rightarrow (P \xrightarrow{\alpha\%} Q).$$

Example of a grammar rule formalization taken from [Herrera et al., 2024].

A first research question concerns the very notion of grammatical description. We wish to draw inspiration from the extensive body of work in symbolic AI and (philosophical) logic devoted to the representation of reasoning and uncertainty [Halpern, 2005], following [Dönicke et al., 2020] for linguistics. The modeling of conditionals (statements of the form “if A , then B ”) and of norms is a long-standing topic in logic, where it is used to model so-called “common sense” reasoning. This reasoning is said to be “non-monotonic”: a conclusion can be called into question in light of new information. For example, in English grammar, one considers that “the subject precedes the verb”, but some constructions invert the subject and force us to revise this rule. Rather than trying to produce grammar rules that describe a corpus completely and exactly, we aim to account for a complex corpus by means of hierarchical, contextual and dynamic rules that incorporate uncertainty and probabilities of occurrence. These rules may come in different families (different logics) depending on the desired level of detail. From this perspective, non-monotonic logics are not only useful for representing common sense reasoning: they also provide suitable foundations for some scientific theories.

A second research question concerns the extraction of these rules from data. Previous work has used machine learning methods for this purpose [Chaudhary et al., 2020, Chaudhary et al., 2022, Herrera et al., 2024, Herrera et al., 2025].

Unfortunately, these approaches are unsatisfactory, notably because they produce a large set of rules that are hard to interpret. In particular, these approaches do not incorporate *a priori* knowledge that could guide the search for patterns. **We will therefore focus on developing algorithms specific to our problem.**

We are looking for a candidate for an end-of-studies internship (Master 2, final year of engineering school), **with the possibility to continue with a PhD** (funding already secured).

How to apply. The candidate will ideally have a background in computer science, with an interest in linguistics. A background in linguistics is a plus, but is not required to apply. We also consider applications from candidates with a more linguistics-oriented background, provided they can demonstrate strong computer science skills. Please send us a CV together with your transcripts for Licence 3, Master 1 and Master 2 (or equivalent), as well as a short description of your interests regarding this internship. There is no need to include a cover letter.

- caio.corro@irisa.fr
- peggy.cellier@irisa.fr
- alexandre.termier@irisa.fr

Applications will be reviewed on a rolling basis until the position is filled.

References

- [Chaudhary et al., 2020] Chaudhary, A., Anastasopoulos, A., Pratapa, A., Mortensen, D. R., Sheikh, Z., Tsvetkov, Y., and Neubig, G. (2020). Automatic extraction of rules governing morphological agreement. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5212–5236, Online. Association for Computational Linguistics.
- [Chaudhary et al., 2022] Chaudhary, A., Sheikh, Z., Mortensen, D. R., Anastasopoulos, A., and Neubig, G. (2022). Autolex: An automatic framework for linguistic exploration.
- [Dönicke et al., 2020] Dönicke, T., Yu, X., and Kuhn, J. (2020). Real-valued logics for typological universals: Framework and application. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3990–4003, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- [Gerdes et al., 2018] Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018). SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In de Marneffe, M.-C., Lynn, T., and Schuster, S., editors, *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.
- [Halpern, 2005] Halpern, J. Y. (2005). *Reasoning about Uncertainty*. Reasoning about Uncertainty. MIT Press, London, England.
- [Herrera et al., 2024] Herrera, S., Corro, C., and Kahane, S. (2024). Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15114–15125, Torino, Italia. ELRA and ICCL.
- [Herrera et al., 2025] Herrera, S., Silai, I.-M., Corro, C., Guillaume, B., and Kahane, S. (2025). Extraction of contrastive rules from syntactic treebanks: A case study in Romance languages. In Chen, X. and Wang, Y., editors, *Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025)*, pages 26–38, Ljubljana, Slovenia. Association for Computational Linguistics.

[Nivre et al., 2020] Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.