

Stage de fin d'étude — IRISA (Rennes)

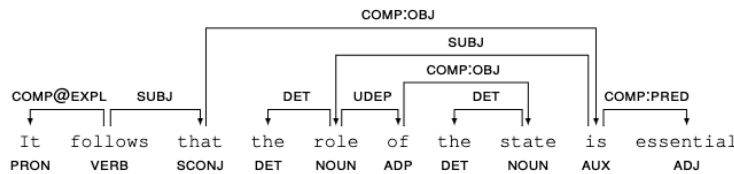
Méthodes informatiques pour la recherche en linguistique formelle

Année universitaire 2026-2027

- Type de contrat : stage de fin d'étude (4-6 mois)
- Laboratoire : IRISA, Rennes
- Équipes de recherche : Archimédia et Lacodam
- Encadrement :
Caio Corro, Peggy Cellier et Alexandre Termier
- Possibilité de continuer sur une thèse financée



Alors que le traitement automatique des langues a longtemps reposé sur une collaboration étroite entre informatique et linguistique, notamment par l'intégration de connaissances linguistiques pour améliorer les modèles, les avancées récentes du domaine se sont faites sans aucun recours à la linguistique. Le projet Autogramm AI propose de repenser cette interdisciplinarité en se donnant pour objectif de faire avancer la recherche en linguistique formelle en utilisant les outils de l'informatique, que ce soit par des méthodes formelles ou des méthodes d'apprentissage automatique, à l'image de l'approche « *AI for science* » qui connaît une grande popularité dans des domaines comme la physique ou le climat. Nous nous intéressons en particulier à l'induction de grammaires descriptives à partir de données linguistiques annotées [Herrera et al., 2024].



Exemple de phrase annotée en syntaxe. Des données suivant ce format d'annotation sont disponibles pour de nombreuses langues via le projet Universal Dependencies [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:

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

Exemple de formalisation de règle de grammaire tiré de [Herrera et al., 2024].

Une première question de recherche concerne la notion même de description grammaticale. Nous souhaitons nous inspirer des nombreux travaux en IA symbolique et en logique (philosophique) consacrés à la représentation du raisonnement et de l'incertitude [Halpern, 2005], dans la lignée de [Dönicke et al., 2020] pour la linguistique. La modélisation des conditionnels (énoncés du type « si A , alors B ») et des normes est un objet d'étude ancien en logique, où elle sert à modéliser le raisonnement dit « de sens commun ». Ce raisonnement est dit « non monotone » : une conclusion peut être remise en cause à la lumière d'une information nouvelle. Par exemple, pour la grammaire de l'anglais, on considère que « le sujet précède le verbe », mais certaines constructions inversent le sujet et obligent à réviser cette règle. Plutôt que de chercher à produire des règles de grammaire décrivant un corpus de manière complète et exacte, nous visons à rendre compte d'un corpus complexe au moyen de règles hiérarchisées, contextuelles et dynamiques, intégrant l'incertitude et des probabilités d'occurrence. Ces règles pourront être déclinées en différentes familles (différentes logiques) selon le niveau de détail souhaité. Dans cette perspective, les logiques non monotones ne servent pas seulement à représenter le raisonnement de sens commun : elles offrent aussi des fondements adaptés à certaines théories scientifiques.

Une seconde question de recherche concerne l'extraction de ces règles à partir de données. Des travaux précédents ont utilisé pour cela des méthodes issues de l'apprentissage automatique [Chaudhary et al., 2020, Chaudhary et al., 2022, Herrera et al., 2024, Herrera et al., 2025]. Malheureusement, ces approches sont peu satisfaisantes, notamment car

elle produisent un grand ensemble de règles difficilement interprétables. En particulier, ces approches n'intègrent pas de connaissances *a priori* qui peuvent guider la recherche de motifs. **Nous nous intéresserons donc à développer des algorithmes spécifiques à notre problématique.**

Nous cherchons un candidat pour stage de fin d'étude (Master 2, dernière année d'école d'ingénieur), **avec une possibilité de continuer sur une thèse** (financement déjà obtenu).

Pour candidater. Le candidat aura idéalement une formation en informatique, avec un intérêt pour la linguistique. Une formation en linguistique est un plus, mais n'est pas requise pour candidater. Nous considérons aussi les candidatures avec une formation plus orientée linguistique, sous condition de pouvoir démontrer une aisance importante en informatique. Merci de nous envoyer un CV ainsi que vos relevés de notes de Licence 3, Master 1 et Master 2 (ou équivalent), ainsi qu'une courte description de vos centres d'intérêt pour ce stage. Il n'est pas nécessaire de joindre une lettre de motivation.

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

Les candidatures seront traitées au fil de l'eau jusqu'à ce que le stage soit pourvu.

Références

- [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.