Bilan Scientifique et Financier
Linglexnumériricorps – Amorçage
Mary C. Lavissière – coordinatrice du projet

Résumé

L'objectif du projet Linglexnumériricops était de commencer à travailler à la création d'un corpus annoté de documents juridiques pouvant être utilisé à des fins pédagogiques. Le projet a atteint cet objectif de plusieurs manières. Tout d'abord, il a permis la consolidation d'un consortium interdisciplinaire d'experts composé de 4 laboratoires et de 9 experts. Ces chercheur.es se sont engagé.es à poursuivre le travail sur un projet ANR. Deuxièmement, le projet a également permis d'identifier le type d'annotation qui serait intéressant pour un corpus de textes juridiques. Il s'agit des *moves* [1] et le type de document qui devrait être annoté, les jurisprudences. Troisièmement, le projet a permis à un corpus pilote de documents juridiques (contrats et jurisprudence) d'être annoté pour les thèmes et les *moves* par des étudiant.es en Master 1 de droit. Quatrièmement, ce projet a permis de rédiger une première version d'une pré-proposition pour une candidature à l'AAP générique de l'ANR. Cette première version sera améliorée en 2022, à l'aide d'un financement supplémentaire demandé sous forme de Maturation afin de soumettre une pré-proposition pour l'AAPG 2023. Cinquièmement, le projet a permis de consolider le terrain par un événement avec les étudiants et les anciens du Master Juriste Trilingue, au cours duquel une présentation du projet a eu lieu et un questionnaire préliminaire sur l'analyse des besoins a été réalisé. Il faut ajouter que le projet a été affecté par le COVID-19, qui a rendu les déplacements impossibles au printemps 2021. L'incertitude du semestre d'automne et les contraintes personnelles (santé et visite de la famille après une longue période d'enfermement) n'ont pas permis à l'équipe de se réunir pleinement en personne, mais les visioconférences et la réunion en personne de l'équipe à Nantes ont permis au projet d'aboutir à la plupart de ses objectifs.

1. Consortium

1) Centre de Recherche sur les Identités, les Nations et l'Interculturalité (CRINI). LSP et équipe de traduction spécialisée. La coordinatrice scientifique de Linglexnumériricorps était Mary C. Lavissière, MCF, spécialisée en LSP anglais et espagnol. Elle codirige l'équipe de recherche LSP. Elle codirige également le Master Juriste Trilingue de l'Université de Nantes. Le projet inclut Johannes Dahm, MCF, co-directeur de l'équipe LSP, qui utilise dans ses recherches des méthodes quantitatives d'analyse du discours et de LSP.

2) Analyse et Traitement Informatique de la Langue Française (ATILF). Alex Boulton, PU et directeur de l'ATILF est une référence internationale en matière de *data-driven learning* (DDL) - éditeur en chef de ReCALL, la revue de l'Association européenne pour l'apprentissage des langues, assisté par ordinateur (EUROCALL) - vice-président de l'Association française de linguistique appliquée (AFLA). L'ATILF est particulièrement adaptée au projet *Linglexnumériricorps* car

elle est spécialisée en linguistique, en didactique des LSP et en informatique. Elle a déjà intégré des projets interdisciplinaires en linguistique et en informatique : Équipex - Ortolang - CNRTL - EDOlang.
3) Laboratoire des Sciences du Numérique de Nantes (LS2N). Équipe Traitement du langage naturel (TAL). Nicolas Hernandez, Christine Jacquin, Laura Monceaux, tous MCF, et Richard Dufour, PU. L'équipe a travaillé sur des projets portant sur l'annotation de corpus (ANR HORAE), l'analyse de discours (FUI ODISAE) et le TAL pour l'éducation (ANR PASTEL). Leur expertise couvre l'apprentissage automatique et profond, la fouille de textes et l'analyse des dialogues. L'équipe est titulaire de la chaire UNESCO, IA-REL (Intelligence artificielle pour l'éducation et les ressources éducatives libres).
4) Laboratoire inter-universitaire de recherche en didactique des langues (LAIRDIL), spécialisé dans la recherche sur la didactique des LSP. Laura Hartwell, PU, professeur titulaire et directrice du LAIRDIL, est une experte de la didactique des LSP et des caractéristiques linguistiques de l'anglais juridique.

2. Type d'annotation et type de document

De nombreuses réunions sur le projet ont été centrées sur l'identification de la question de recherche qui sous-tend le projet en linguistique, didactique et sciences du numérique du type d'annotation qui serait intéressant pour les textes juridiques. Ces unités ont été identifiées en linguistique comme des *moves* par [1] C'est à ce point que la recherche actuelle en sciences du numérique est à sa limite. L'identification d'autres types d'unités plus petites a déjà été réalisée par d'autres chercheurs [2,3]. Cependant, les unités de discours plus grandes n'ont pas été systématiquement étudiées dans la jurisprudence [4] Au cours de ces réunions, il est apparu clairement que les questions de recherche les plus urgentes en informatique sont centrées sur l'identification des parties plus grandes du discours [5, 6]. Enfin, si la littérature cite l'utilisation de ces divisions à des fins pédagogiques [7], aucune étude ne vérifie actuellement l'efficacité de ces divisions dans l'apprentissage des langues. Ces lacunes seront affinées au cours de l'année 2022 afin de proposer leur investigation pour un projet ANR dans l'AAPG2023.

3. Corpus pilote

Au cours du projet, un corpus pilote a été annoté par les étudiants de M1 du Master Juriste Trilingue. Ils ont utilisé le logiciel Webanno [8], qui permet à plusieurs annotateurs de travailler sur le même projet. Il dispose également d'une interface relativement simple. Les étudiant.es ont annoté un contrat pour sa macrostructure et une jurisprudence pour ses *moves*. Les étudiants ont constaté que le logiciel n'était pas adapté à l'annotation de plus grandes unités de discours car il était nécessaire d'annoter les documents ligne par ligne. Un nouveau logiciel sera utilisé dans la prochaine phase du projet. Ce logiciel doit permettre d'annoter de plus grandes unités à la fois. Une autre difficulté était le manque de littérature décrivant la nature exacte des *moves*. Une définition claire des *moves* sera

entreprise lors de la prochaine étape du projet. Enfin, le logiciel Webanno ne permet pas de mesurer la variation des annotations lorsque les annotations sont créées par le chercheur et non celles qui font partie des annotations incluses dans le logiciel. Un nouveau logiciel qui inclut une mesure de la variation entre annotateurs sera utilisé pour la prochaine étape du projet. Amorçage a donc permis d'affiner la méthodologie du futur projet.

4. Dépôt d'un projet pour l'AAPG 2022 de l'ANR

Amorçage a permis au consortium de déposer une candidature à l'AAPG 2022 de l'ANR. Ce projet est intitulé *Lexhnology: joint linguistic and NLP discourse structure modeling of legal texts for language pedagogy*. Il a été sélectionné par le comité scientifique pour la seconde étape du processus de sélection de l'AAPG 2022.

5. Événement de communication et analyse des besoins

Le financement d'Amorçage a permis l'organisation d'un événement pour faire connaître le projet sur le terrain et pour commencer une analyse préliminaire des besoins. Cet événement a eu lieu à l'Université de Nantes. Trois promotions du Master Juriste Trilingue ont été invitées ainsi que les 30 diplômé.es du programme. Environs soixante personnes ont participé à l'événement, qui a eu lieu le 26 novembre 2021. Les participant.es ont répondu à un questionnaire, qui a été pré-testé en ligne en avril 2021.

Bilan financier en pièce jointe.

References

[1] J. Swales, *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press, 2004. doi: 10.1017/CBO9781139524827.
[2] D. Hoadley, Blackstone. ICLR&D, 2019. https://github.com/ICLRandD/Blackstone
[3] LexNLP. 2018. https://github.com/LexPredict/lexpredict-lexnlp
[4] S. Gozdz-Roszkowski, “Move Analysis of Legal Justifications in Constitutional Tribunal Judgments in Poland: What They Share and What They Do Not,” *Int J Semiot Law*, vol. 33, no. 3, pp. 581–600, Sep. 2020, doi: 10.1007/s11196-020-09700-1.
[5] V. Araujo, A. Villa, M. Mendoza, M.-F. Moens, and A. Soto, “Augmenting BERT-style Models with Predictive Coding to Improve Discourse-level Representations,” In EMNLP, Nov. 2021.
[6] L. Huber, C. Memmadi, M. Dargnat, and Y. Toussaint. Do sentence embeddings capture discourse properties of sentences from scientific abstracts ? In the First ACL Workshop on Computational Approaches to Discourse, pages 86–95, 2020.
[7] E. Cotos, S. Huffman, and S. Link, “A move/step model for methods sections: Demonstrating Rigour and Credibility,” *English for Specific Purposes*, vol. 46, pp. 90–106, Apr. 2017, doi: 10.1016/j.esp.2017.01.001.
[8] S.M. Yimam, R. Eckart de Castilho, I. Gurevych, and C. Biemann “Automatic Annotation Suggestions and Custom Annotation Layers in WebAnn,” In: Proceedings of ACL-2014, demo session, Baltimore, 2014.