



Séminaire du CeRLA



Centre de Recherche en Linguistique Appliquée



VENDREDI 23 JANVIER 2026


salle **CLI.036** campus BDR

Nicolas Ballier (Université Paris Cité)

9:30-12:00 Masterclass sur les agents conversationnels NotebookLm et LMstudio

Cette masterclass s'attachera à présenter deux outils en un sens complémentaires. Notebook LM, qui permet de s'initier à une intelligence artificielle qui est explicite dans ses sources (Grounded AI) et qui participe aux problématiques de répliquabilité et de traçabilité. Comment, à partir d'un document, poser des questions à un modèle de langue (Gemini 2.5) en ayant l'assurance que les réponses seront fondées exclusivement sur le document ou les documents fournis. C'est le principe de ce qu'on appelle le RAG (Retrieval Augmented Generation). Cette application de Google, qui pour le moment autorise une certaine liberté (car elle n'a pas encore défini son modèle économique) propose un certain nombre de fonctionnalités que je commenterai. La plus connue est la génération de podcasts, c'est la plus utilisée, y compris à des fins pédagogiques. Je proposerai rapidement quelques exploitations de recherche tant sur les travaux de l'analyse de la factualité que sur les problématiques associées aux voix ou à ce que l'on nomme les personae des agents conversationnels.

Le deuxième type d'agent conversationnel de l'IA générative que je souhaiterais présenter est l'interface avec des modèles locaux, et constitue ce que l'on pourrait appeler une initiation à l'IA frugale. A la différence du premier modèle, on est ici assuré de la souveraineté de ses données, les fichiers qui servent à l'analyse restent notre propriété, sont stockés sur notre ordinateur ainsi que les réponses des modèles. J'exposerai les trois principales fonctionnalités ou les trois profils principaux d'utilisation de ces modèles et présenterai quelques utilisations possibles pour l'enseignement et éventuellement pour la recherche.

 Pour profiter pleinement de l'atelier, les participant-e-s sont invité-e-s à venir avec leur **ordinateur** portable et à **installer** à l'avance LMstudio (<https://lmstudio.ai/>) qui nécessite un compte gmail, ainsi qu'un modèle au choix (en fonction de la mémoire disponible sur l'ordinateur), par exemple le modèle gpt-oss-20b.

14:00-16:00 Présentation: « Se positionner comme linguiste dans des recherches sur l'IA : l'exemple des biais de tokenisation »

Cette présentation détaillera des travaux sur l'analyse des biais de genre dans les systèmes de traduction et des travaux en cours sur les grands modèles de langues, et plus spécifiquement les modèles de langues utilisés pour la parole comme Whisper. J'expliquerai quelques méthodes utilisées pour sonder les modèles neuronaux et les principaux résultats obtenus. La réflexion portera sur ce que les linguistes peuvent apporter à l'analyse de systèmes jugés pour la qualité de leur performance. On fera retour sur ces modalités d'analyse de ces systèmes (benchmarks) et sur la place que les linguistes peuvent y prendre. Je contribue, avec d'autres, à examiner ce que deviennent les langues et la connaissance que

nous pouvons en avoir dans ce type de représentation. En particulier, j'insisterai sur l'étape dite de "tokenisation" qui permet de préciser ce que voit le réseau de neurones, le type d'entités manipulées par la machine. Je terminerai en esquisant la transformation en cours de sciences du langage en science des données linguistiques et ce qu'elle implique en partie.

Bibliographie sélective

Wisniewski, G. Zhu, L. Ballier, N. and Yvon, F. (2022) Analyzing Gender Translation Errors to Identify Information Flows between the Encoder and Decoder of an NMT System, *BlackboxNLP2022, EMNLP2022*,

Nicolas Ballier, N., Arnold, T., Méli, A, Thurston, T. & Yunès, J.B (2024) Whisper for L2 speech scoring, *The International Journal of Speech Technology*. Springer,

Ballier, N & Méli, A. (2024) Investigating Acoustic Correlates of Whisper Scoring for L2 Speech Using Forced alignment with the Italian Component of the ISLE corpus, *NLP4CALL 2024, Rennes, 24-25 Oct 2024*, 20-32.

Ballier, N., Burin, L., Namdarzadeh, B., Ng, S, Wright, R. and Yunès, J.-B. (2024) Probing Whisper Predictions for French, English and Persian Transcriptions, *7th International Conference on Natural Language and Speech Processing, October 19-20, 2024, Trento, Italy*, 129-138,

Liang, S., Ballier, N., Levow, G. A., & Wright, R. (2025). Beyond WER: Probing Whisper's Sub-token Decoder Across Diverse Language Resource Levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 31225-31235).