Paul Lerner

Email : lerner@isir.upmc.fr
LinkedIn : paul-lerner-466997a8
GitHub : github.com/PaulLerner
Site web : https://paullerner.github.io/

Études

Qualification MCF section 27 (Informatique)

Qualification pour candidater aux concours de Maître de conférences (MCF) dans la section 27 (Informatique) du CNU, valable jusqu'en 2028 inclus. Rapporteurs : Sylvain Castagnos (Université de Lorraine) et Thierry Delot (Université Polytechnique Hauts-de-France).

Université Paris-Saclay, CNRS, LISN (ex-LIMSI) Orsay, France
Thèse en Informatique, Répondre aux questions visuelles à propos d'entités nommées 2020–2023

Thèse dirigée par Olivier Ferret (Université Paris-Saclay, CEA, List) et co-encadrée par Camille Guinaudeau (Université Paris-Saclay, CNRS, LISN).

Membres du jury :

Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LISN)	Président
Josiane Mothe (IRIT, CNRS, Université Toulouse Jean-Jaurès)	Rapporteur & Examinatrice
Philippe Mulhem (Université Grenoble Alpes, CNRS, LIG)	Rapporteur & Examinateur
Michel Crucianu (CEDRIC-CNAM)	Examinateur
Ewa Kijak (Université de Rennes, Inria, IRISA)	Examinatrice

Écoles d'été suivies dans ce cadre :

— ALPS 2021 : traitement automatique des langues et de la parole
— LxMLS 2023 : apprentissage automatique, en particulier appliqué au traitement des langues

Université Paris Descartes Paris, France
Master Sciences, Technologies, Santé, spécialité Intelligence Artificielle, mention bien 2018–2019

Master préparé dans le cadre d'un double diplôme avec l'ESILV.

Stage de fin d'études : *Diagnostic de la maladie de Parkinson d'après un examen manuscrit*, encadré par Laurence Likforman-Sulem (Télécom ParisTech).

ESILV Courbevoie, France
Ingénieur en Informatique 2014–2019

Stage effectué en 1re année d'école (M1) : *Stratégies comportementales pendant une interaction homme-machine*, encadré par Beatrice Biancardi et Catherine Pelachaud (ISIR, Sorbonne Université).

Czech Technical University Prague, République Tchèque
Licence en informatique (échange avec l'ESILV) 2016

Expériences

Sorbonne Université, CNRS, ISIR Paris, France
Chercheur postdoctoral – Biais politiques des grands modèles de langues 2024–2026

Postdoctorat encadré par François Yvon et Benjamin Piwowarski au sein de l'équipe MLIA (*Machine Learning for Information Access*) de l'ISIR.

Mots-clés : résumé automatique, traduction automatique, grand modèle de langue multilingue, biais politiques, alignement, multiculturalisme.

Sorbonne Université, CNRS, ISIR Paris, France
Chercheur postdoctoral – Traduction automatique des néologismes scientifiques 2023–2024

Postdoctorat encadré par François Yvon au sein de l'équipe MLIA de l'ISIR.

Mots-clés : néologisme, variation terminologique, morphologie, traduction automatique, grand modèle de langue multilingue, *in-context learning*.

Université Paris-Saclay, CNRS, LISN (ex-LIMSI) Orsay, France
Doctorant – Répondre aux questions visuelles à propos d'entités nommées 2020–2023

Thèse dirigée par Olivier Ferret (Université Paris-Saclay, CEA, List) et co-encadrée par Camille Guinaudeau (Université Paris-Saclay, CNRS, LISN), au sein de l'équipe Traitement du Langage Parlé (TLP) du LISN.

Mots-clés : questions visuelles, recherche d'information multimodale, apprentissage de représentation, entités nommées, pré-entraînement, système de question-réponse

Université Paris-Saclay, CNRS, LIMSI Orsay, France
Ingénieur de recherche – Identification multimodale du locuteur 2019–2020

Travail encadré par Hervé Bredin et Camille Guinaudeau, au sein de l'équipe Traitement du Langage Parlé (TLP) du LIMSI.

Mots-clés : dialogue multipartites, identification du locuteur, reconnaissance d'entités nommées, désambiguïsation d'entités nommées, alignement forcé, apprentissage actif, multimodalité

Télécom ParisTech Paris, France
Stagiaire – Diagnostic de la maladie de Parkinson d'après un examen manuscrit Mars–Septembre 2019

Stage encadré par Laurence Likforman-Sulem au sein de l'équipe Signal, Statistique et Apprentissage (S2A) de Télécom ParisTech.

Mots-clés : maladie de Parkinson, apprentissage de représentation, aide au diagnostic

Sorbonne Université, CNRS, ISIR Paris, France
Stagiaire – Apprentissage par renforcement pour une stratégie comportementale Avril–Septembre 2018

Stage encadré par Beatrice Biancardi et Catherine Pelachaud au sein de l'équipe PIRoS à l'ISIR.

Mots-clés : interaction humain-machine, apprentissage par renforcement, agent conversationnel

Encadrement

Joanna Radoła – Génération d'alternance codique Paris, France
Stage M2 – Sorbonne Université, CNRS, ISIR Février–Août 2025

Stage co-encadré par François Yvon.

Mots-clés : alternance codique, grand modèle de langue multilingue

Salem Messoud – Réordonnancement multimodal Orsay, France
Stage M2 – Université Paris-Saclay, CNRS, LISN Mars–Septembre 2022

Stage co-encadré par Olivier Ferret et Camille Guinaudeau.

Mots-clés : recherche d'information multimodale, réordonnancement, système de question-réponse, questions visuelles, apprentissage de représentation, entités nommées

Enseignement

Enseignant à Aivancity (vacataire) Cachan, France
Natural Language Processing 2024–2025

Année	Matière	Établissement	Niveau	CM	TD	TP
2024-2025	*Natural Language Processing*	Aivancity	M1 et M2*	16	–	32
2024-2025	*Deep Learning : Models and Optimization*	ENSAE	M2 (Ingé 3)	–	–	6
2024-2025	*Machine Learning for Natural Language Processing*	ENSAE	M2 (Ingé 3)	–	–	9
2023-2024	*Deep Learning : Models and Optimization*	ENSAE	M2 (Ingé 3)	–	–	6
2023-2024	*Machine Learning for Natural Language Processing*	ENSAE	M2 (Ingé 3)	–	–	9
2022-2023	Introduction à l'apprentissage statistique	UFR Sciences Paris-Saclay	L3	–	–	24
2022-2023	Bases du développement logiciel	Polytech Paris-Saclay	L3 (Ingé 1)	–	24	–
2022-2023	Programmation impérative	Polytech Paris-Saclay	L1 (Prépa 1)	–	14	10
2021-2022	Introduction à l'apprentissage statistique	UFR Sciences Paris-Saclay	L3	–	–	24
2021-2022	Bases du développement logiciel	Polytech Paris-Saclay	L3 (Ingé 1)	–	24	–
2021-2022	Programmation impérative	Polytech Paris-Saclay	L1 (Prépa 1)	–	14	10
2020-2021	Bases du développement logiciel	Polytech Paris-Saclay	L3 (Ingé 1)	–	22	–
2020-2021	Programmation impérative	Polytech Paris-Saclay	L1 (Prépa 1)	–	16	16

Total éq. TD : 284h

**Le même cours a été donné au 1er semestre aux M1 et au 2e semestre aux M2*

Conception de l'intégralité des cours magistraux (CM) et travaux pratiques (TP). Conception et correction d'un devoir maison et d'un examen par semestre. Support : https://paullerner.github.io/aivancity_nlp/

Notions abordées : loi de Zipf, sémantique distributionnelle, skip-gram, plongements lexicaux, réseaux de neurones récurrents, mécanisme d'attention, Transformer, grands modèles de langues, préentraînement, ajustement (*fine-tuning*), alignement (RLHF/DPO), décodage/génération, *in-context learning*, métriques d'évaluation, éthique et biais.

Chargé de TP à l'ENSAE (vacataire) — Palaiseau, France

Deep Learning : Models and Optimization — 2023–2025

Responsable : Kevin Scaman (2023-2024) puis Olivier Koch (2024-2025). Support : https://kscaman.github.io/teaching/2023_ENSAE_DL.html

Notions abordées : perceptron multi-couches, classification multi-classes (entropie croisée), réseau de neurones convolutionnel, traitement d'images, auto-encodeur variationnel.

Chargé de TP à l'ENSAE (vacataire) — Palaiseau, France

Machine Learning for Natural Language Processing — 2023–2025

Encadrement de projets et correction de rapports de projets. Responsable : Christopher Kermorvant. Support : https://github.com/Deep-NLP-Course

Notions abordées : Sac de mots et classification, plongements lexicaux (*word embedding*) et analogies, modèles de langues et génération.

Chargé de TD à l'UFR Sciences, Université Paris-Saclay (moniteur) — Orsay, France

Introduction à l'apprentissage statistique — 2021–2023

Introduction à l'apprentissage automatique et au traitement automatique des langues. Encadrement et soutenance de projets ainsi que surveillance et correction d'examens. Responsables : François Landes et Kim Gerdes. Support : https://gitlab.inria.fr/flandes/ias

Notions abordées : Descente de gradient, Perceptron, Analyse en composantes principales, Estimateur du maximum de vraisemblance, Classification naïve bayésienne, K-moyennes, TF-IDF, Surapprentissage et généralisation, Pré-traitement et encodage.

Chargé de TD à Polytech Paris-Saclay (moniteur) — Orsay, France

Programmation Impérative — 2020–2023

Surveillance et correction de TP notés et de devoirs maisons ainsi que surveillance d'examens. Responsable : Frédéric Voisin.

Notions abordées : Représentations binaires, types, boucles et conditions, fonctions, algorithmes, récursivité.

Chargé de TD à Polytech Paris-Saclay (moniteur) Orsay, France
Bases du développement logiciel 2020–2023

Surveillance et correction de TP notés et de devoirs maisons ainsi que surveillance d'examens. Responsable : Joël Falcou.

Notions abordées : Bibliothèque standard du C++, flux entrée/sortie, pointeurs et références, surcharge d'opérateurs et de fonctions, struct, template, tests unitaires.

Responsabilités collectives

Relecteur

Année	Revue	Conférence internationale	Conférence nationale
2024		EMNLP, ACMMM, ACL, ICMR	JEP-TALN
2023	Pattern Recognition	ICMR	RECITAL-RJCRI
2022		ACMMM	

Représentant des CDD au conseil du laboratoire Orsay, France
Université Paris-Saclay, CNRS, LIMSI Octobre 2019–2020

Participation aux réunions mensuelles du conseil.

Publications

Revues internationales avec comité de lecture

[1] B. Biancardi, M. Mancini, **P. Lerner** et C. Pelachaud, « Managing an Agent's Self-Presentational Strategies During an Interaction », *Frontiers in Robotics and AI*, t. 6, p. 16, 2019, **Impact factor : 3.4**, Long, ISSN : 2296-9144. DOI : 10.3389/frobt.2019.00093.

Revues nationales avec comité de lecture

[2] **P. Lerner**, S. Messoud, O. Ferret, C. Guinaudeau, H. Le Borgne, R. Besançon, J. G. Moreno et J. Lovón Melgarejo, « Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées », *Traitement Automatique des Langues*, t. 63, n° 2, p. 15-39, 2022, Long. adresse : https://aclanthology.org/2022.tal-2.0.

Conférences internationales avec comité de lecture

[3] **P. Lerner** et F. Yvon, « Towards the Machine Translation of Scientific Neologisms », in *Proceedings of the 31st International Conference on Computational Linguistics*, **Rang B**, Long, 9 pages, International Committee on Computational Linguistics, 2025. adresse : https://aclanthology.org/2025.coling-main.63/.

[4] **P. Lerner** et F. Yvon, « Unlike "Likely", "Unlike" is Unlikely : BPE-based Segmentation hurts Morphological Derivations in LLMs », in *Proceedings of the 31st International Conference on Computational Linguistics*, **Rang B**, Court, 5 pages, International Committee on Computational Linguistics, 2025. adresse : https://aclanthology.org/2025.coling-main.348/.

[5] **P. Lerner**, O. Ferret et C. Guinaudeau, « Cross-modal Retrieval for Knowledge-based Visual Question Answering », in *Advances in Information Retrieval (ECIR 2024)*, **Rang A**, Long, Présentation, Cham : Springer Nature Switzerland, 2024, p. 421-438, ISBN : 978-3-031-56027-9. DOI : 10.1007/978-3-031-56027-9_26.

[6] **P. Lerner**, O. Ferret et C. Guinaudeau, « Multimodal Inverse Cloze Task for Knowledge-Based Visual Question Answering », in *Advances in Information Retrieval (ECIR 2023)*, **Rang A**, Long, Présentation, Cham : Springer Nature Switzerland, 2023, p. 569-587, ISBN : 978-3-031-28244-7. DOI : 10.1007/978-3-031-28244-7_36.

[7] **P. Lerner**, J. Bergoënd, C. Guinaudeau, H. Bredin, B. Maurice, S. Lefevre, M. Bouteiller, A. Berhe, L. Galmant, R. Yin et C. Barras, « Bazinga ! A Dataset for Multi-Party Dialogues Structuring », in *Proceedings of the Language Resources and Evaluation Conference*, **Rang B**, Long, Poster, Marseille, France : European Language Resources Association, 2022, p. 3434-3441. adresse : https://aclanthology.org/2022.lrec-1.367.

[8] **P. Lerner**, O. Ferret, C. Guinaudeau, H. Le Borgne, R. Besançon, J. G. Moreno et J. Lovón Melgarejo, « ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities », in *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, sér. SIGIR'22, **Rang A***, Long, Poster, New York, NY, USA : Association for Computing Machinery, 2022, p. 3108-3120. DOI : 10.1145/3477495.3531753. adresse : https://hal.science/hal-03650618/.

[9] M. Mancini, B. Biancardi, S. Dermouche, **P. Lerner** et C. Pelachaud, « Managing Agent's Impression Based on User's Engagement Detection », in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, sér. IVA '19, **Rang B**, Court, Poster, 3 pages, Paris, France : Association for Computing Machinery, 2019, p. 209-211, ISBN : 9781450366724. DOI : 10.1145/3308532.3329442.

Conférences nationales avec comité de lecture

[10] **P. Lerner** et F. Yvon, « Vers la traduction automatique des néologismes scientifiques », French, in *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, M. Balaguer, N. Bendahman, L.-M. Ho-dac, J. Mauclair, J. G Moreno et J. Pinquier, éd., **Rang C**, Long, Présentation, Toulouse, France : ATALA et AFPC, juill. 2024, p. 245-261. adresse : https://aclanthology.org/2024.jeptalnrecital-taln.17.

[11] **P. Lerner**, O. Ferret et C. Guinaudeau, « Recherche cross-modale pour répondre à des questions visuelles », in *18e Conférence en Recherche d'Information et Applications*, H. Zargayouna, éd., **Rang C**, Long, Présentation, Paris, France : ATALA, 2023, p. 74-92. adresse : https://aclanthology.org/2023.jeptalnrecital-coria.5.

[12] **P. Lerner**, O. Ferret, C. Guinaudeau, H. Le Borgne, R. Besançon, J. G. Moreno et J. Lovón Melgarejo, « Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées en utilisant des bases de connaissances », in *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 2022.*, **Rang C**, Court, Présentation, Avignon, France : ATALA, 2022, p. 434-444. adresse : https://aclanthology.org/2022.jeptalnrecital-taln.43/.

[13] **P. Lerner** et L. Likforman-Sulem, « Classification of Online Handwriting Time Series for Parkinson's Disease Diagnosis using Deep Learning », in *Proceedings of the 4th Junior Conference on Data Science and Engineering (JDSE) (non-archival)*, Court, Poster, 2019, p. 3.

Ateliers internationaux avec comité de lecture

[14] **P. Lerner** et C. Grouin, « INCLURE : a Dataset and Toolkit for Inclusive French Translation », in *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, P. Zweigenbaum, R. Rapp et S. Sharoff, éd., Long, Présentation, 10 pages, Torino, Italia : ELRA et ICCL, mai 2024, p. 59-68. adresse : https://aclanthology.org/2024.bucc-1.7.

Séminaires invités

[15] **P. Lerner**, « Morphological Competence of LLMs : Applied to Translation of Scientific Neologisms », in *ChangeLing (CMU) seminars*, online, 2025. adresse : https://changelinglab.github.io/.

[16] **P. Lerner**, « Automatic Data Annotation and Webly Supervised Visual Question Answering », in *Knowledge-Enhanced Information Retrieval workshop (KEIR @ ECIR 2024)*, Glasgow, Scotland, 2024. adresse : https://keirworkshop.github.io/.

[17] **P. Lerner**, « Towards Machine Translation of Scientific Neologisms », in *IRIT seminars*, Toulouse, France, 2024. adresse : https://www.irit.fr/EVT/PDF/evt-1070-en.pdf.

Productions logicielles et de données

[18] **P. Lerner**, *Inclure*, version v0.1.0, 7 avr. 2024. adresse : https://github.com/PaulLerner/inclure.

[19] **P. Lerner**, *neott*, version v1.2.0, 17 déc. 2024. adresse : https://github.com/PaulLerner/neott.

[20] **P. Lerner**, *ViQuAE*, version v4.0.0-alpha, 16 jan. 2024. adresse : https://paullerner.github.io/ViQuAE/.

Programmation

- **Apprentissage de représentation :** PyTorch
- **Langages :** Python, C, C++, C#, Java
- **Bibliothèques Python :** Faiss, Transformers, spaCy, Scikit-learn, NumPy, Matplotlib, Seaborn
- **Clusters :** slurm
- **Bases de données :** SPARQL, Elasticsearch

Langues

- **Français :** langue maternelle
- **Anglais :** courant
- **Espagnol :** scolaire