

Alice MILLOUR

Maîtresse de Conférences Section 27

UNIVERSITÉ PARIS 8 VINCENNES SAINT-DENIS

Laboratoire LIASD – Équipe PASTIS

UFR STN – Licence IV

2 774, route du Col de Porte

38700 Le Sappey-en-Chartreuse

+33 6 35 57 88 67

am@up8.edu

<https://alicemillour.github.io>

Mots-clé : Linguistique computationnelle, sciences participatives, évaluation (données et modèles), diversité linguistique, langues non standardisées.

PARCOURS ACADÉMIQUE

- Depuis sept. 2022 **Maîtresse de Conférences en informatique**, UFR STN – Licence IV, Laboratoire LIASD – Équipe PASTIS.
UNIVERSITÉ PARIS 8 VINCENNES SAINT-DENIS, Saint-Denis.
- Janv. - août 2022 **Post-doctorante**, Étiquetage morphosyntaxique et lemmatisation automatiques pour la langue Corse (6 mois), UMR LISA.
CNRS ET UNIVERSITÀ DI CORSICA, Corte.
- 2019 - 2021 **Attachée Temporaire d'Enseignement et de Recherche (ATER)** à temps plein (deux CDD d'un an chacun).
SORBONNE UNIVERSITÉ, Paris.

PARCOURS UNIVERSITAIRE

- 2016 - 2020 **Doctorat** mention *mathématiques, informatique et applications*
Laboratoire « Sens, Texte, Informatique, Histoire », SORBONNE UNIVERSITÉ, Paris

« Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées »
- 2015 - 2016 **Master 2 recherche**, Traitement automatique des langues (TAL), SORBONNE PARIS IV
« Construction de ressources langagières par myriadisation (crowdsourcing) pour le traitement automatique des langues peu dotées, le cas de l'Alsacien »
- 2011 - 2015 **École d'ingénieur en informatique**, Filière Ingénierie des Systèmes d'Information, ENSIMAG, Grenoble

ENSEIGNEMENTS

Les volumes sont renseignés en heures équivalent TD (HETD).

2022 - 2024 :

MCF à Paris 8, Licence Informatique & Vidéoludisme

Intitulé	Public	Année	Volume
Méthodologie de la programmation	L1	2022-24	TD : 45 / CM : 67.5
Pratique des machines	L1	2022-24	TD : 30 / CM : 45
Outils informatiques collaboratifs	L1	2023-24	TD : 22.5 / CM : 33.75
Algorithmique et structures de données 2	L2	2022-23	TD : 22.5 / CM : 33.75
Ingénierie des langues	L3	2022-24	TD : 43.75 / CM : 22.5

2019 - 2021 :

ATER à SORBONNE UNIVERSITÉ, Master Sciences du langage : Langue et Informatique

Intitulé	Public	Année	Volume
Théorie des langages (Grammaires formelles et stochastiques)	M1	2019-20	TD : 9.75 / CM : 9.75
Modèles de linguistique computationnelle	M1	2020-21	TD : 9 / CM : 10.5
Programmation de modèles linguistiques II (sémantique)	L3	2019-21	TD : 24 / CM : 36
Bases de données	L3	2019-21	TD : 52 / CM : 4.5
Mathématiques et statistiques (logiciel R)	L3	2020-21	TD : 19.5
Ateliers professionnels sciences du Langage	L1	2019-21	TD : 39
Préparation au PIX (ancien C2i : tableur, présentations, outils collaboratifs)	L1/L2/L3	2019-21	TD : 144 / CM : 27

Supports de cours produits (voir : <https://alicemillour.github.io/teaching/>)

Deux des cours de bases de données dispensés avec K. Fort ont été largement inspirés des activités « débranchées » *bases de données à tricoter* et *enquête de police* proposées par Marie Duflot-Kremer (voir : <https://members.loria.fr/MDuflot/files/med/index.html>).

RESPONSABILITÉS UNIVERSITAIRES

Au sein de l'université Paris 8 Vincennes Saint-Denis.

2023 - 2024 Référente violences sexuelles et sexistes (UFR STN)

2023 - 2024 Responsable de la L1 Informatique et Vidéoludisme (90 étudiant.e.s)

ENCADREMENT DE THÈSE

2023 - Marina Seghier : **Explicabilité des modèles de TAL par les ressources : caractérisation de la variation textuelle et paradigme d'évaluation systématique**, co-encadrement avec Nicolas Jouandeau.

PUBLICATIONS (DERNIÈRE MÀJ. 2023)

Lorsqu'il est disponible, le classement CORE 44¹ (pour les conférences) ou SJR 45² (pour les revues) est précisé.

Revues nationales avec comité de lecture

La Revue TAL est la revue principale en France pour le traitement automatique des langues.

- Alice Millour and Karën Fort. À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. In *Revue TAL : numéro spécial sur les langues peu dotées (59-3)*. Association pour le Traitement Automatique des Langues, 2018. 25p. Disponible sur : <https://www.atala.org/content/traitement-automatique-des-langues-peu-dot%C3%A9es>, Revue SJR "Q3/Q4"

¹Voir : <http://portal.core.edu.au/conf-ranks/>.

²Voir : <http://www.scimagojr.com/>.

Conférences internationales avec comité de lecture

- Alice Millour and Karën Fort. Toward a Lightweight Solution for Less-resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing. In *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai 2018. 6p. Disponible sur <http://www.lrec-conf.org/proceedings/lrec2018/pdf/326.pdf>, CORE "C"
- Alice Millour and Karën Fort. Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling. In *Recent Advances in Natural Language Processing (RANLP)*, pages 776 – 784, Varna, Bulgarie, septembre 2019. Poster. Disponible sur <https://www.aclweb.org/anthology/R19-1090/>, CORE "C"
- Lionel Nicolas, Verena Lyding, Claudia Borg, Corina Forascu, Karën Fort, Katerina Zdravkova, Iztok Kosem, Jaka Čibej, Špela Arhar Holdt, Alice Millour, Alexander König, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Anisia Katinskaia, Anabela Barreiro, Lavinia Aparaschivei, and Yaakov HaCohen-Kerner. Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Actes de the 12th Language Resources and Evaluation Conference*, pages 268–278, Marseille, France, mai 2020. European Language Resources Association. Disponible sur <https://www.aclweb.org/anthology/2020.lrec-1.34/>
- Alice Millour. Getting to Know the Speakers: a Survey of a Non-Standardized Language Digital Use. In *Actes de 9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2019)*, Poznan, Pologne, mai 2019. 6p. Présentation. Voir : <http://ltc.amu.edu.pl/>
- Alice Millour, Marianne Grace Araneta, Ivana Lazic Konjik, Annalisa Raffone, Yann-Alan Pilatte, and Karën Fort. Katana and Grand Guru: a Game of the Lost Words (DEMO). In *Actes de 9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2019)*, Poznan, Pologne, mai 2019. 2p. Poster et démonstration. Voir : <http://ltc.amu.edu.pl/>

Workshops internationaux avec comité de lecture

- Alice Millour and Karën Fort. Text Corpora and the Challenge of Newly Written Languages. In *Actes de 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 111–120, Marseille, France, mai 2020. European Language Resources association. Disponible sur : <https://www.aclweb.org/anthology/2020.sltu-1.15/>
- Alice Millour and Karën Fort. Krik: First steps into crowdsourcing pos tags for kréyòl gwadloupéyen. In *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Workshop CCURL, Miyazaki, Japon, mai 2018. 6p. Disponible sur http://lrec-conf.org/workshops/lrec2018/W26/pdf/10_W26.pdf
- Alice Millour and Karën Fort. Why do we Need Games? Analysis of the Participation on a Crowdsourcing Annotation Platform. In *Actes de Games4NLP 2017 - Using Games and Gamification for Natural Language Processing*, Symposium Games4NLP, Valence, Espagne, avril 2017. 2p. Voir <https://anawiki.essex.ac.uk/dali/games4nlp17/#presentations>

Conférences nationales avec comité de lecture

- Alice Millour, Karën Fort, Delphine Bernhard, and Lucie Steible. Vers une solution légère de production de données pour le TAL : création d'un tagger de l'alsacien par crowdsourcing

bénévole. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, pages 139–154, Orléans, France, juin 2017. Disponible sur <https://www.aclweb.org/anthology/2017.jeptalnrecital-long.10/>. CORE "C"

Atelier national avec comité de lecture

- Laurent Kevers and Alice Millour. Réalisations, obstacles et perspectives pour l'outillage du corse. In *Actes des Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, Marseille, France, 11 2022. CNRS
- Yoann Dupont, Carlos-Emiliano González-Gallardo, Gaël Lejeune, Alice Millour, and Jean-Baptiste Tanguy. QUEER@DEFT2021 : Identification du profil clinique de patients et notation automatique de copies d'étudiants (QUEER@DEFT2021 : Patients clinical profile identification and automatic student grading). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, pages 95–107, Lille, France, 6 2021. ATALA
- Alice Millour, Karën Fort, and Pierre Magistry. Répliquer et étendre pour l'alsacien "étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux". In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, pages 29–37, Nancy, France, juin 2020. Présentation, (en ligne) . Disponible sur <https://www.aclweb.org/anthology/2020.jeptalnrecital-eternal.4/>

Séminaires invité

- Alice Millour. Table ronde d'ouverture. In *Contribuling, Inalco*, Paris, France, 2023
- Alice Millour. Production participative et ressources linguistiques, quels enjeux pour la diversité linguistique en TAL ? In *Séminaire Cognition & Langage de Maxime Amblard et Manuel Rebuschi*, IDMC (Institut des Sciences du Digital - Management & Cognition), Nancy, février 2019. Voir : <https://idmc.univ-lorraine.fr/informations-master-sciences-cognitives/info-sc-seminaires-archives/>
- Alice Millour. Construction d'un corpus annoté en parties du discours par myriadisation (crowd-sourcing) pour le traitement automatique d'une langue peu dotée : l'alsacien. In *séminaire « Recherches linguistiques et corpus » organisé par Franck Neveu*, Sorbonne Université, Paris, avril 2017

Conférences internationales sans comité de lecture

- Alice Millour and Karën Fort. Production participative de ressources linguistiques variées pour l'alsacien. In *NWADV'2019 (New Ways of Analyzing Dialectal Variation)*, Paris, France, novembre 2019. Présentation, voir <https://sites.google.com/view/nwadv2019>
- Alice Millour, Marianne Araneta, Ivana Lazic Konjik, Annalisa Raffone, Yann-Alan Pilatte, and Karën Fort. Quest game or "Katana and Grand Guru: A Game of the Lost Words". In *enetCollect Cost Action (CA16105) 3rd Annual Meeting*, Lisbonne, Portugal, mars 2019. Poster, voir : https://www.enetcollect.net/ilias/goto.php?target=file_829_download, présentation et démonstration, voir : https://www.enetcollect.net/ilias/goto.php?target=file_715_download

Conférences nationales sans comité de lecture

- [Alice Millour](#) and Karën Fort. Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling. In *Journées inaugurales du GDR LIFT*, Orléans, France, novembre 2019. Poster. Voir <https://gdr-lift.loria.fr/journees-scientifiques-de-lancement-du-gdr-orleans-28-29-novembre-2019/>

Ouvrages de vulgarisation

- [Alice Millour](#) and Karën Fort. Sciences participatives et diversité linguistique, Retours d'expériences. *Culture et recherche*, 140 (hiver 2019-2020):91–92, décembre 2019. Disponible sur : <https://www.culture.gouv.fr/content/download/264287/3004796?version=5>
- [Alice Millour](#) and Karën Fort. Recettes de grammaire, sciences participatives et langues non standardisées. *Culture et recherche*, 140 (hiver 2019-2020):92, décembre 2019. 1p. Disponible sur : <https://www.culture.gouv.fr/content/download/264287/3004796?version=5>

CO-ENCADREMENT DE MÉMOIRE DE MASTER

- | | |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2023 | Liam Duignan (M1 - Université Paris Cité) : Paradigmes d'évaluation interprétables pour la tâche de reconnaissance des entités nommées en français , avec Jean-Yves Antoine. |
| 2023 | Marina Seghier (M2 - Sorbonne Nouvelle Paris 3, INALCO, Paris Nanterre) : Caractérisation objective des domaines/genres/registres/modalités pour le TAL : le cas de la reconnaissance des entités nommées en français , avec Jean-Yves Antoine. |
| 2022 | Yang Heng (M1 - Université de Strasbourg) : Détection de la variation graphique dans une langue non standardisée (dialectes alsaciens) , avec Delphine Bernhard et Pablo Ruiz Fabo. |
| 2021 | Alexane Jouglar (M1 - SU) : Évaluation d'outils d'annotation en entités nommées pour le français : que compare-t-on ? , avec Karën Fort et Yoann Dupont. |
| 2019 | Harmonie Begue (M1 - SU) : Développement de ressources langagières et d'outils de TAL pour le créole mauricien , avec Karën Fort et Gaël Lejeune. |
| 2017 | Gwladys Feler (M1 - SU) : Entraînement d'un analyseur morphosyntaxique pour le créole guadeloupéen , avec Karën Fort (non soutenu). |

AUTRES ACTIVITÉS LIÉES À LA RECHERCHE

Participation à des projets de recherche

- | | |
|-------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2023 - | (À venir) Membre de l'ANR DiviTAL |
| 2019 - | Co-responsable avec Karën Fort du groupe de réflexion « Variations » au sein du GDR LIFT |
| 2016 - 2020 | Groupe de travail « <i>Crowdsourcing</i> implicite pour la production de matériel pédagogique pour l'enseignement des langues » de l'action COST <i>European Network for Combining Language Learning with Crowdsourcing Techniques</i> — ENETCOLLECT. |
| 2017 - 2018 | Projet « Langues et numérique » (Ministère de la culture) PLURAL (Production LUdique de Ressources Annotées pour les Langues de France), en collaboration avec B. Guillaume (Loria), K. Fort (Sorbonne Université), D. Bernhard (LiLPa) et A. Thibault (Sorbonne Université). |

Collaborations internationales

- 2020 Membre de la tâche 2 (« How do we use crowdsourcing to help maintain or revitalize endangered languages? ») du Crowdfest organisé par ENETCOLLECT à Coimbra, Portugal, du 2 au 4 février.
- 2019 Membre du groupe 1 (« Quest game ») du Hackaton organisé par ENETCOLLECT à Bruxelles, Belgique, du 22 au 25 janvier. Développement d'un prototype de jeu vidéo encourageant la transmission et l'apprentissage de langues non standardisées pâtissant du phénomène de transfert linguistique.
- 2018 Séjour de recherche (*Short Term Scientific Mission*) à l'Université du Pays-Basque UPV/EHU, San Sebastián, Espagne, en collaboration avec R. Agerri (Équipe IXA-NLP), dans le cadre d'ENETCOLLECT (voir ci-dessous) du 12 au 17 février. Élaboration d'un sondage interrogeant les pratiques du *crowdsourcing* en Europe.

Organisation de séminaires et colloques

- 2020 Co-organisatrice, avec Gaël Lejeune et Melchior Simioni de la "Journée d'étude SIBON: Sociologie et Informatique", 16 janvier 2020, Sorbonne Université
- 2019 Co-organisatrice, avec Mathieu Avanzi et André Thibault du colloque international *New Ways of Analyzing Dialectal Variation* (NWADV'2019), 21-23 novembre 2019, Paris, France

Organisation d'ateliers

- 2017 Co-organisatrice de l'atelier **TALN DiLiTAL** (Diversité Linguistique et TAL), 2017, Orléans, France.

Participation à des comités de lecture

Lorsqu'il est disponible, le classement CORE 44³ (pour les conférences) ou SJR 45⁴ (pour les revues) est précisé. Le nombre d'articles est donné entre parenthèses.

Revues

- Italian philosophy journal *Etica&Politica/Ethics&Politics* - 2023 (1), Revue TAL (numéro thématique "la robustesse des systèmes de traitement automatique des langues", 2023) (1) ;
- Relectrice de deuxième niveau : *Information Processing and Management* 2019 (CORE "A") (1).

Conférences et ateliers internationaux

- Games4NLP'2024 (Atelier de LREC'2024) (1), LREC'2024 (Core "B") (3), EMNLP 2023 (4), EACL 2022 (2), COLING 2022 (3), Games 2022 (2), COLING 2020 (CORE "A") (1), ACL-IJCNLP 2020 (CORE "A") (1), AAAI 2020 (CORE "A*") (1), LREC'2020 (CORE "C") (2), Games4NLP'2020 (Atelier de LREC'2020) (1), LRL'19 (Atelier de LTC 2019) (1) ;
- Relectrice de deuxième niveau : ACL 2019 (CORE "A*") (1), The Web Conference 2019 (CORE "C") (1), NAACL 2019 (CORE "A") (1).

Conférences et ateliers nationaux

- Robustal 2022, TALN 2022, RECITAL 2019 (Jeunes chercheurs) (2) ;
- Relectrice de deuxième niveau : TALN 2017 (CORE "C") (1).

³Voir : <http://portal.core.edu.au/conf-ranks/>.

⁴Voir : <http://www.scimagojr.com/>.

Événements grand public / médiation / valorisation

- 2018 Conférence « Langues et numérique 2018 » (3 juillet), Délégation générale à la langue française et aux langues de France (DGLFLF) — Présentation du projet PLURAL.
- 2018 Conférence de presse pour le lancement de la plateforme RECETTES DE GRAMMAIRE (11 juin), dans les locaux de l'OLCA, Strasbourg.
- 2016 et 2017 Fête de la science (octobre), Paris, France — Présentation de ZOMBILINGO.

Intégration dans la communauté

- 2020 - 2021 Membre du comité éditorial des actes de la conférence *New Ways of Analyzing Dialectal Variation*, en collaboration avec M. Avanzi et A. Thibault, collection « Travaux de linguistique et de philologie » (TraLiPhi) de la Société de linguistique romane
- 2019 Étudiante bénévole pour RANLP'2019 (vérification de la conformité des articles, des plannings, des actes)
- 2018 - 2019 Coorganisation avec G. Lejeune du « **Dojo de Deep Learning** » hebdomadaire à Sorbonne Université.
- 2018 Présentation au **séminaire** mensuel « Linguistique Computationnelle » (17 avril) : *Crowdsourcing POS tags for Kréyòl Gwadeloupéyen*, Sorbonne Université.
- 2016 Étudiante bénévole durant JEP-TALN-RECITAL 2016

BOURSES / PRIX

- Bourse **ATALA** — TALN 2017.
- Bourse du **SIAL** (Service Interuniversitaire pour l'Apprentissage des Langues), Sorbonne Université — EACL 2017.

OUTILS ET RESSOURCES LANGAGIÈRES DÉVELOPPÉS

« Myriadisation » est un terme qui a été proposé par Gilles Adda comme traduction française du terme anglais *crowdsourcing*.

Plateformes de myriadisation

Plateformes libres de droit pour la production participative de **corpus bruts, annotés en parties du discours**, et de **lexiques de variantes dialectales et orthographiques**. Les codes source sont publiés sur GITHUB, à l'adresse <https://github.com/alicemillour/Bisame>.

Ressources langagières myriadisées

Les corpus sont annotés avec le jeu d'étiquettes universel (Voir : universaldependencies.org/) complété pour les besoins de chaque langue. Les ressources produites sur les plateformes de myriadisation pour l'alsacien, le créole guadeloupéen et le créole mauricien sont respectivement disponibles sur la plateforme Ortolang aux adresses https://www.ortolang.fr/market/corpora/bisame_gsw/v1, https://www.ortolang.fr/market/item/krik_gcf et https://www.ortolang.fr/market/item/ayo_mfe-creole-mauricien.

Prototype de jeu ayant un but

Katana and Grand Guru: a Game of the Lost Words (voir <https://bisame.paris-sorbonne.fr/lost-words/>, via Firefox) est un jeu vidéo qui se joue en collaboration entre un apprenant et un locuteur d'une langue donnée. L'apprenant doit, pour progresser dans le jeu, réaliser des tâches linguistiques qui permettent d'encourager la transmission linguistique inter-générationnelle tout en produisant la collecte de ressources langagières. .

Voir publications éponymes ainsi que la documentation du projet réalisé dans le cadre des Crowdfest ENETCOLLECT sur le dépôt GitHub : [alicemillour/KatanaAndGG](https://github.com/alicemillour/KatanaAndGG).