

Alice MILLOUR

Née le 13 septembre 1991 en France
Trilingue français - anglais - espagnol
italien (niveau C1), allemand (A2)

60 bis rue Haxo, 75020, Paris
+33 6 35 57 88 67
alice.millour@etu.sorbonne-universite.fr
<https://alicemillour.github.io>

Doctorante contractuelle

Linguistique computationnelle, sciences participatives, diversité linguistique

EXPÉRIENCES PROFESSIONNELLES

- ÉTÉ 2016 **Développeuse web** pour ZOMBILINGO, Nancy (stage de deux mois).
Ajout d'une fonctionnalité pour le *crowdsourcing* d'annotation linguistique en morpho-syntaxe au sein d'une plateforme existante : ZOMBILINGO.
- 2014 - 2015 **Responsable d'application** chez ATOS WORLDLINE, Lyon (stage de huit mois).
Projets MÉGALIS BRETAGNE et E-BOURGOGNE : mise en place de la solution d'authentification unique (OPENAM), développements pour une application web (JAVA), automatisation des scripts de déploiement (PYTHON).

FORMATION

- DEPUIS 2016 **Doctorat** (3^e année), mention *mathématiques, informatique et applications* sous la direction de Karën Fort et Claude Montacé :
« *Construction de ressources langagières par myriadisation (crowdsourcing) pour le traitement automatique des langues peu dotées : le cas des langues de France.* ». Laboratoire STIH (Sens, Texte, Informatique, Histoire) - ED5 (Concepts et langage) - SORBONNE UNIVERSITÉ, Paris.
- 2015 - 2016 **Master 2 recherche en Traitement automatique des langues (TAL)** : « *Construction de ressources langagières par myriadisation (crowdsourcing) pour le traitement automatique des langues peu dotées, le cas de l'Alsacien.* ». SORBONNE PARIS IV, Paris.
- 2011 - 2015 **Diplôme d'ingénieur en informatique**. ENSIMAG, Grenoble.
Filière : ISI (Ingénierie des Systèmes d'Information)
- 2013 - 2014 **Échange universitaire à l'étranger**. UTFSM, Valparaíso, Chili.
- 2008 - 2010 **Classe préparatoire aux concours des grandes écoles** option physique-chimie. LYCÉE SAINT-LOUIS, Paris.
- 2008 **Baccalauréat S - S.V.T.** (mention très bien). Lycée Jules Ferry, Paris .

BOURSES / PRIX

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

OUTILS ET RESSOURCES LANGAGIÈRES CRÉÉS

Plateformes de myriadisation

Plateformes libres de droit pour la production participative de **corpus bruts et annotés en parties du discours**, et de **lexiques de variantes dialectales et orthographiques**.

Les plateformes BISAME, KRIK et RECETTES DE GRAMMAIRE ont été déployées pour l'alsacien et le créole guadeloupéen et sont adaptables à d'autres langues facilement (voir la documentation sur le dépôt GitHub : [alicemillour/Bisame](https://github.com/alicemillour/Bisame)).

Corpus annotés en morphosyntaxe

Les corpus sont annotés avec le jeu d'étiquettes universel [Petrov *et al.*, 2012] complété pour les besoins de chaque langue.

Alsacien	corpus de 7 973 tokens annoté par les participants de BISAME (F-mesure : 0,93) [Millour et Fort, 2018c].
Créole guadeloupéen	corpus de référence de 933 tokens annoté avec la participation de G. Feler et A. Thibault [Millour et Fort, 2018b].

CO-ENCRADREMENT DE MÉMOIRES DE MASTER 1

- 2019 Harmonie Begue : **Développement de ressources langagières et d'outils de TAL pour le créole mauricien**, avec Karën Fort et Gaël Lejeune, SORBONNE UNIVERSITÉ, Paris.
- 2017 Gwladys Feler : **Entraînement d'un analyseur morphosyntaxique pour le créole guadeloupéen**, avec Karën Fort, SORBONNE UNIVERSITÉ, Paris (non soutenu).

PUBLICATIONS

Revue nationale avec comité de lecture

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

- Millour, A. et Fort, K. (2018a). 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. Revue SJR "Q3/Q4"

Conférences internationales avec comité de lecture

- Millour, A. et Fort, K. (2018c). 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. CORE "C"

Workshops internationaux avec comité de lecture

- Millour, A. et Fort, K. (2018b). Krik: First steps into crowdsourcing pos tags for kréyòl gwadeloupéyen. *In Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Workshop CCURL, Miyazaki, Japon
- Millour, A. et Fort, K. (2017). 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

Conférences nationales avec comité de lecture

- Millour, A., Fort, K., Bernhard, D. et Steible, L. (2017). 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)*, Orléans, France. CORE "C"

Séminaires invité

- Millour, A. (2019). 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
- Millour, A. (2017). Construction d'un corpus annoté en parties du discours par myriadisation (crowdsourcing) pour le traitement automatique d'une langue peu dotée : l'alsacien. *In séminaire « Recherches linguistiques et corpus » organisé par Franck Neveu*, Sorbonne Université, Paris

Conférences internationales sans comité de lecture

- Araneta, M., Fort, K., Lazi Konjik, I., Millour, A., Pilatte, Y.-A. et Raffone, A. (2019). Quest game or "Katana and Grand Guru: A Game of the Lost Words". *In enetCollect Cost Action (CA16105) 3rd Annual Meeting*, Lisbonne, Portugal. (présentation orale, poster et démonstration)

AUTRES ACTIVITÉS LIÉES À LA RECHERCHE

Collaborations internationales

- 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.

Participation à des projets de recherche

- 2016 - 2020 Groupe de travail « *Crowdsourcing* implicite pour la production de matériel pédagogique pour l'enseignement des langues » de l'action COST *European Network for Combining Language Learning with Crowdsourcing Techniques* – 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).

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 44¹ (pour les conférences) ou SJR 45² (pour les revues) est précisé. Le nombre d'articles est donné entre parenthèses.

Revues internationales

- Relectrice de deuxième niveau : Information Processing and Management 2019 (CORE "A") (1 article).

Conférences et ateliers internationaux

- 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 nationales

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

Médiation et événements grand public

- 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

- 2018 - 2019 Participation au « **Dojo Deep Learning** » hebdomadaire organisé par G. Lejeune à Sorbonne Université.
- 2018 Présentation au **séminaire d'équipe** mensuel « Linguistique Computationnelle » (17 avril) : *Crowdsourcing POS tags for Kréyòl Gwadeloupéyen*, Sorbonne Université.
- 2016 Étudiante bénévole durant JEP-TALN-RECITAL 2016

COMPÉTENCES INFORMATIQUES

Langages maîtrisés :	PYTHON, JAVASCRIPT, PHP, MYSQL, HTML/CSS, BASH, \LaTeX .
Autres langages connus :	C, ASSEMBLEUR, JAVA, ADA.
IDE :	Emacs, vim, Eclipse, NetBeans.
Outils de TAL :	AntConc, Gate, entraînement de pos-taggers MEMM – MELT [Denis et Sagot, 2012] Deep learning – Framework KERAS
Outils de gestion :	Git, TortoiseHg, Maven, Jenkins.

¹Voir : <http://portal.core.edu.au/>.

²Voir : <http://www.scimagojr.com/>.

