Ingénierie linguistique

Linguistique descriptive, linguistique de corpus, traitement automatique des langues

Marine Delaborde

marine.delaborde@gmail.com
marine-delaborde
lattice-cnrs.com

Cursus

14/12/2020 • **Doctorat** en sciences du langage — Université Sorbonne Nouvelle - Paris III
Cf. Parcours professionnel.

2013 • **Master** en ingénierie linguistique (PluriTAL) — Université Sorbonne Nouvelle - Paris III
M2 : Parcours « Recherche et développement », mention bien.
- Mémoire : analyse automatique de citations dans des articles scientifiques.
- Directrice de mémoire : Isabelle Tellier

2011 • **Licence** en sciences du langage — Université Sorbonne Nouvelle - Paris III
Parcours général, mention assez bien.

Formation

- **Informatique** (Master PluriTAL) :
  * **OS** : Linux (Ubuntu), Mac, Windows.
  * **Bureautique** : Open Office, iWork, Microsoft Office, LaTeX.
  * **Programmation** : Python, Bash, Perl, R, HTML, XML, JavaScript.
  * **Bases de données** : MySQL, MongoDB.
- **Linguistique - Informatique** (Formation continue, master, thèse) :
  * **Annotation et analyses** : Analec, TXM, CLAN, ELAN, Praat.
  * **Textométrie** : TXM, iTrameur, Lexico.
  * **Corpus** : Frantext (CQL).
  * **Automates** : Unitex, Nooj.
  * **TAL et ingénierie linguistique** : NLTK, Bonsai, Wordnet, WOLF, Weka, Wapiti.
  * **Suivi de l'école d'été ESSLLI** (logique, langage, information) en 2017.
- **Langues** : anglais (lu, écrit, parlé), espagnol (notions).
- **Divers** :
  * **Sauveteur Secouriste du Travail** depuis 2018 (Recyclage effectué en janvier 2020).
  * **Diplôme de moniteur en éducation canine** de la Société Centrale Canine (2018).
  * **Stage de spécialisation "école des chiots"** de la Société Centrale Canine (2018).

Parcours professionnel

Novembre 2016 - Décembre 2020 • **Thèse** en sciences du langage Université Sorbonne Nouvelle - Paris III (ED 622)
Contrat doctoral avec l'École Normale Supérieure (du 01/11/2016 au 31/10/2019).
- Sujet : analyse en corpus de chaînes de coréférence.
- Directeur de thèse : Frédéric Landragin
- Laboratoire : Lattice (UMR 8094)
- Projet : DEMOCRAT (cf. section page 3).
- Productions scientifiques : 1 à 7 (cf. section page 4).

Mars - Octobre 2016 • **Ingénieure d'études** *(CDD temps plein)* Université de Versailles Saint-Quentin-en-Yvelines
Responsables : Jean-Claude Yon (CHCSC - UVSQ) & Anne Vilnat (LIMSI - CNRS)
Projet PRESNUM (cf. section page 3). Analyse textuelle de critiques musicales et cinématographiques via l'extraction automatique d'informations.

Mai 2014 - Décembre 2015
- **Ingénieure d'études** *(CDD temps plein)* LIMSI - CNRS
  Responsable : Patrick Paroubek
  Projet SONAR (cf. section page 3) : extraction automatique d'informations dans des offres d'emploi.

Mars - Juillet 2013
- **Stagiaire** *(Stage de M2 - temps plein)* LIMSI - CNRS
  Responsable : Patrick Paroubek
  Projet MECALIT (cf. section page 3) + analyses d'articles ISCA : réalisation d'une chaîne de traitements automatiques pour l'analyse de citations dans des articles scientifiques en mécanique des fluides et en traitement de la parole. Productions scientifiques : 8 et 9 (cf. section page 4).

Juillet 2011
- **Vacataire** *(CDD temps plein)* Laboratoire de phonétique et phonologie, Paris III
  Responsable : Cécile Fougeron et Cédric Gendrot
  Rangement et référencement informatique des ouvrages de la bibliothèque du laboratoire.

Enseignement

octobre 2020 - août 2022
- **ATER** en TAL à l'ILPGA - Université Sorbonne Nouvelle
  - « **La coréférence** ». Public : Master plurital, 1ère année.
  - « **Apprendre à programmer avec python** ». Public : L3 et Master.
  - « **Linguistique de corpus** ». Public : Licence de sciences du langage, 3ème année.
  - « **Informatique et industrie de la langue** ». Public : Licence de sciences du langage, 2ème année.
  - « **Programmation pour les humanités numériques** ». L3 et Master.
  - « **Statistiques textuelles** ». L3 et Master.

2018 - 2019 Semestre 2
- **Chargée de cours** *(Vacataire)* Université Sorbonne Nouvelle - Paris III
  - UFR : ILPGA. « **Informatique et industrie de la langue** ». Public : Licence de Sciences du Langage, 2 groupes de 18 et 19 étudiants de 2ème année. Responsable : Serge Fleury. Contenu : Outillage informatique pour la linguistique. Projet de récupération automatique de corpus sur le Web et traitements textométriques. Durée : 39h (TD) + 1h30 (CM).

2018 - 2019 Semestre 1
- **Chargée de cours** *(Vacataire)* Université Sorbonne Nouvelle - Paris III
  - UFR : Arts et Médias. « **Humanités numériques** ». Public : Licence de Médiation culturelle, 63 étudiants de 2/3ème années. Responsable : Kim Gerdes. Contenu : Google Ngram Viewer, Frantext (CQL), expressions régulières, formats de fichiers, documents structurés. Durée : 7,5h.
  - UFR : BET. « **Les humanités numériques** ». Public : Licence - Enseignements Transversaux, 26 étudiants de toutes les années. Contenu : Google Ngram Viewer, Frantext (CQL), expressions régulières, formats de fichiers, documents structurés. Responsable : Kim Gerdes. Durée : 7,5h.
  - UFR : ILPGA. « **Introduction aux humanités numériques** ». Public : Licence de Sciences du Langage, 51 étudiants de 1ère année en Mineure Humanités Numériques. Contenu : cours d'introduction et dernier cours de révisions (fouille de texte, expressions régulières, XML). Responsable : Ioana Galleron. Durée : 4h.

2017 - 2018 Semestre 2
- **Chargée de cours** *(Vacataire)* Université Sorbonne Nouvelle - Paris III
  - UFR : ILPGA. « **Informatique et industrie de la langue** ». Public : Licence de Sciences du Langage, 2 groupes de 14 et 19 étudiants de 2ème année. Responsable : Serge Fleury. Contenu : Outillage informatique pour la linguistique. Projet de récupération automatique de corpus sur le Web et traitements textométriques. Durée : 39h (TD) + 1h30 (CM).

2017 - 2018
Semestre 1

- **Chargée de cours** *(Vacataire)* Université Sorbonne Nouvelle - Paris III
  - UFR : Arts et Médias. « **Humanités numériques** ». Public : Licence de Médiation culturelle, 73 étudiants de 2/3ème années. Responsable : Kim Gerdes. Contenu : Google Ngram Viewer, Frantext (requêtes avancées), formats de fichiers. Durée : 7,5h.
  - UFR : BET. « **Les humanités numériques** ». Public : Licence - Enseignements Transversaux, 38 étudiants de toutes les années. Contenu : Frantext (requêtes avancées), formats de fichiers. Responsable : Kim Gerdes. Durée : 3,5h.

2016 - 2017
Semestre 2

- **Chargée de cours** *(Vacataire)* Université Sorbonne Nouvelle - Paris III
  - UFR : ILPGA. « **Informatique et industrie de la langue** ». Public : Licence de Sciences du Langage, 2 groupes de 19 et 20 étudiants de 2ème année. Responsable : Serge Fleury. Contenu : Outillage informatique pour la linguistique. Projet de récupération automatique de corpus sur le Web et traitements textométriques. Durée : 39h (TD) + 1h30 (CM).

Participation à des projets

2016 - 2020

- **DEMOCRAT** : Projet ANR regroupant 4 laboratoires : Lattice, LiLPa, ICAR et IHRIM. Sur le thème des chaînes de référence en linguistique descriptive, linguistique de corpus et traitement automatique des langues. Participation aux journées annuelles du projet (communications), présentation de poster (production scientifique 6), annotation de textes en coréférence, description de phénomènes linguistiques, rédaction de recommandations pour des systèmes de TAL.

Mars - Octobre 2016

- **PRESNUM** : Collaboration entre le CHCSC - UVSQ et le LIMSI - CNRS pour une analyse textuelle de critiques musicales et cinématographiques. Réalisation de chaînes de traitements impliquant du pré-traitement de données textuelles (nettoyage, formatage : données structurées), de l'extraction automatique d'informations (détection des champs : auteurs, notes, commentaires) et des statistiques textuelles (richesse de vocabulaire, longueurs de phrases, recouvrement de vocabulaire, etc.).

Mai 2014 - Décembre 2015

- **SONAR** : Projet FUI. Collaboration entre Multiposting, Work4, LIASD et LIMSI - CNRS. En collaboration étroite avec l'équipe de développement de Multiposting (visites à l'entreprise régulières). Réalisation de chaînes de traitements de données textuelles impliquant de la récupération de corpus (MongoDB), des pré-traitements (nettoyage, formatage de données) et de l'extraction automatique d'informations dans des offres d'emploi (tâches, compétences) à l'aide de patrons linguistiques. Rédaction de règles d'extraction afin de réaliser un corpus d'entraînement.

Mars - Juillet 2013

- **MECALIT** : Projet interne visant la collaboration entre des équipes de mécanique des fluides et de traitement du langage. Réalisation de chaînes de traitements de données textuelles impliquant de la récupération de corpus (aspirations web), des pré-traitements (nettoyage, formatage de données), de l'extraction automatique d'informations dans des articles scientifiques traitant de problèmes de mécanique des fluides (références bibliographiques), et des statistiques textuelles.

Responsabilités collectives

- Décembre 2021 • **Membre du comité d'organisation et du comité scientifique** de la conférence internationale « Entre féminin et masculin – langue(s) et société ».
- 2017 - 2019 • **Représentante élue des doctorants** de l'ED 268 de Paris III . Participation aux conseils et aux bureaux de l'école doctorale ainsi qu'aux auditions pour les bourses de doctorat. Rédaction de comptes rendus.
- 2017 - 2019 • **Représentante suppléante des doctorants** du laboratoire Lattice. Participation au conseil du laboratoire en l'absence du titulaire. Renseignements pour les nouveaux arrivants.
- Mai 2017 • **Bénévole** lors de la conférence NAMED 2017. Aide logistique en amont. Accueil des participants.

Productions scientifiques

1. **Communication - Conférence internationale** : Marine Delaborde, Frédéric Landragin. De la coréférence exacte à la coréférence complexe : une typologie et sa mise en œuvre en corpus. 10èmes Jounées internationales de Linguistique de Corpus, Université Grenoble Alpes, Nov 2019, Grenoble, France. <hal-02286100>
2. **Article de revue** : Marine Delaborde, Frédéric Landragin. En quoi le pronom « on » a-t-il une valeur anaphorique ? Le cas des successions d'occurrences de « on ». Les cahiers de praxématique, Montpellier : Presses universitaires de la Méditerranée, 2006-, 2019, La gestion de l'anaphore en discours : complexités et enjeux, 72, pp.1-18. <hal-02161902>
3. **Communication - Journée d'études** : Frédéric Landragin, Marine Delaborde. Faut-il compter ou ignorer les occurrences de « ce » dans les chaînes de coréférences ? Ce disant, que fait-on ? Aspects grammaticaux et discursifs de ce en français, Université de Strasbourg, 2018, Strasbourg, France. <halshs-01836380>
4. **Communication - Colloque international** : Marine Delaborde, Frédéric Landragin. En quoi le pronom « on » a-t-il une valeur anaphorique ? Le cas des successions d'occurrences de « on ». Gérer L'Anaphore en Discours (GLAD 2018) : vers une approche interdisciplinaire / Managing Anaphora in Discourse : towards an interdisciplinary approach , Apr 2018, Grenoble, France. <halshs-01795213>
5. **Communication - Conférence internationale** : Marine Delaborde, Frédéric Landragin. Traitement « good-enough » du pronom « on » : vers une modélisation de la coréférence floue. Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS) 2018, Jan 2018, Paris, France. <halshs-01795228>
6. **Poster - Conférence nationale** : Frédéric Landragin, Marine Delaborde, Yoann Dupont, Loïc Grobol. Description et modélisation des chaînes de référence. Le projet ANR Democrat (2016-2020) et ses avancées à mi-parcours. Cinquième édition du Salon de l'Innovation en TAL (Traitement Automatique des Langues) et RI (Recherche d'Informations), May 2018, Rennes, France. 2018. <hal-01797982>
7. **Communication - Journée des doctorants du Lattice** : Présentation du sujet de thèse au laboratoire et réponses aux questions d'un discutant défini à l'avance (Jeanne-Marie Debaisieux). 45 minutes en tout.

8. **Article - Conférence internationale** : Joseph Mariani, Christopher Cieri, Gil Francopoulo, Patrick Paroubek, Marine Delaborde : Facing the Identification Problem in Language-Related Scientific Data Analysis. LREC 2014 : 2199-2205
   Lien : http://www.lrec-conf.org/proceedings/lrec2014/pdf/945_Paper.pdf

9. **Article (invité) - Conférence internationale** : Joseph Mariani, Patrick Paroubek, Gil Francopoulo, Marine Delaborde : Rediscovering 25 years of discoveries in spoken language processing : a preliminary ISCA archive analysis. INTERSPEECH 2013 : 3371-3403
   Lien : https://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_3371.pdf

Bénévolat

- **Éducation canine** :

Depuis Mars 2016 - Monitrice bénévole : Club Canin de la Vallée de la Bièvre.
  * Cours d'agility (principalement) - éducation canine - école des chiots.
  * Participation à l'organisation de compétitions nationales et internationales.
  * Participation à des stages et des compétitions nationales et internationales.
  * Communication sur les réseaux sociaux.

2014 - 2015 - Webmaster bénévole : Club Canin de Verrièrres-le-Buisson (site modifié depuis).