

Yoann Dupont

*Docteur en Sciences du Langage,
informaticien de formation,
TAL, apprentissage automatique, génie
logiciel*

France
+33 0 00 00 00 00
✉ yoa.dupont@gmail.com
📄 [yoanndupont.github.io](https://github.com/yoanndupont)
in [yoann-dupont-385b0462](https://www.linkedin.com/in/yoann-dupont-385b0462)
👤 YoannDupont
permis B



Expérience professionnelle

- depuis Nov. **ATER, Sorbonne Université.**
- 2019 Reconnaissance des entités nommées (REN) dans le domaine juridique, REN structurée, REN sur des textes historiques. Encadrement de 4 mémoires dont un sur l'annotation de données libres en en EN et un sur l'étude des biais sociologiques dans les modèles de langues. Participation au jury des M1.
- Juin 2018 **Post-doctorant, Laboratoire d'Informatique Fondamentale d'Orléans (LIFO).**
(1.5 an) Travaux sur les expressions multimots (MWE) au sein de l'ANR PARSEME-FR selon 3 axes : 1. donner une représentation des MWE selon un formalisme de metagrammaires, 2. intégrer leur parsing dans un outil existant qui doit être étendu pour 3. utiliser l'outil sur de larges corpus pour enrichir des lexiques avec de nouvelles MWE.
- Avr. 2018 **Ingénieur d'étude, École Normale Supérieure (ENS).**
(2 mois) Contribution au développement d'un outil de reconnaissance des entités nommées (EN) pour les textes littéraires au sein du département LILA. Encadrement de stagiaires pour adaptation de [SEM](#) sur un corpus de littérature du 18ème au 20ème siècle. Développement d'une interface d'annotation rapide des EN.
- Nov. 2013 **Ingénieur informatique, Expert System France.**
(4 ans) Implémentation de méthodes pour la reconnaissance des entités nommées structurées par apprentissage automatique à l'aide de CRF. Définition des besoins pour l'intégration d'un module d'apprentissage actif. Participation au projet IMM (IRT SystemX) pour intégrer un module d'extraction de relations entre entités nommées : génération de traits pour le module MultiR, basés sur parsing en dépendances syntaxiques.
- Oct. 2012 **Ingénieur d'études, Université Pierre et Marie Curie.**
(3 mois) Intégration d'un module de reconnaissance des entités nommées par apprentissage automatique (CRF). Module intégré dans [SEM](#), logiciel libre open source, de ma création.
- Avr. 2012 **chercheur stagiaire (master), Laboratoire Lattice.**
(6 mois) Combinaisons de méthodes symboliques (généralisation d'automates) et d'apprentissage statistique (CRF) appliquées à la tâche de chunking syntaxique.
- Avr. 2010 **chercheur stagiaire (licence), Laboratoire d'Informatique Fondamentale d'Orléans.**
(3 mois) Intégration de connaissances dans un CRF pour l'annotation morphosyntaxique.

Expérience d'enseignement

- depuis Nov. **ATER, Sorbonne Université.**
- 2019 Enseignements :
- (prévisionnel sur 2 ans)
- cours de M1 modèles de linguistique computationnelle (PoS tagging, évaluation, HMM, distance de Levenstein). Durée : 12h CM et 19.5h TD,
 - cours de L2 introduction au TAL (regex, formats de données, textométrie). Durée : 28,5h CM, 52h TD
 - PIX/C2I. Durée : 13h CM et 174h TD
 - Initiation à la programmation. Durée : 12h TD
- Sep. 2018 **chargé de travaux dirigés (vacataire), INaLCO.**
(1 semestre) Master 2 ingénierie linguistique, module « Langages de script », encadré par Jean-Michel Daube. 1 groupe de 20 étudiants. Durée du TD : 39h (assuré 30h sur 39h). Introduction à Python3, Notebooks, formats de données, TALN, librairies pour l'apprentissage automatique (numpy, scikit-learn, gensim)

- Jan. 2016 **chargé de travaux dirigés (vacataire)**, *Université Sorbonne Nouvelle Paris 3*.
(1 semestre) Licence 2 Sciences du Langage, module « Informatique et Industries de la Langue », encadré par Serge Fleury, Kim Gerdes et Isabelle Tellier. 1 groupe de 20 étudiants. Durée du TD : 18h. Introduction au TALN (google n-grams, frantext, expressions régulières, gromoteur, le Trameur).
- Jan. 2015 **chargé de travaux dirigés (vacataire)**, *Université Sorbonne Nouvelle Paris 3*.
(1 semestre) Licence 2 Sciences du Langage, module « Informatique et Industries de la Langue », encadré par Serge Fleury, Kim Gerdes et Isabelle Tellier. 1 groupe de 20 étudiants. Durée du TD : 18h. Introduction au TALN (google n-grams, frantext, expressions régulières, gromoteur, le Trameur).
- Sept. 2013 **chargé de travaux dirigés (vacataire)**, *Université Sorbonne Nouvelle Paris 3*.
(1 semestre) Licence 2 Sciences du Langage, module « Informatique et Industries de la Langue », encadré par Serge Fleury, Kim Gerdes et Isabelle Tellier. 3 groupes de 10 à 20 étudiants. Durée du TD : 9h.
- Jan. 2012 **Tuteur**, *Université d'Orléans*.
(1 semestre) Tuteur pour les étudiants de première année de licence informatique en difficulté.

Expérience d'administration et autres responsabilités

- 2021 **Relecteur**, *ACL-IJCNLP*.
Relecture de 2 papiers long et 2 papiers courts (papiers courts avec sub-reviewers) dans la track *Resources and Evaluation*.
- 26–30 **Bénévole**, *SyntaxFest*.
Août 2019 Bénévole durant SyntaxFest 2019 à Paris (INaLCO).
- 4–8 Juillet **Bénévole**, *INaLCO*.
2016 Bénévole durant la conférence TALN 2016 à Paris (INaLCO).
- 2015–2017 **représentant des doctorants**, *Université Sorbonne Nouvelle Paris 3*.
Représentant des doctorants de l'école doctorale 268.

Formation

- 2013 – 2017 **Thèse CIFRE**, *Université Sorbonne-Nouvelle Paris 3 et Expert System France*, soutenue le 23 / 11 / 2017.
« La structuration dans les entités nommées ». Extraction de traits pertinents pour améliorer la reconnaissance des entités nommées, développement de techniques pour récupérer la structure interne des entités nommées, comparaison de méthodes par apprentissage (CRF et réseaux de neurones). Développement d'un système état-de-l'art sur le French Treebank en entités nommées.
- 2010 – 2012 **Master d'informatique**, *Université d'Orléans*.
Spécialité IRAD (Informatique, Répartition et Aide à la Décision)
- 2006 – 2010 **Licence d'informatique**, *Université d'Orléans*.
Spécialité STIC (Sciences et Technologies de l'Information et de la Communication)

Compétences

Langages	Python (bibliothèque TAL), Java (module de CRF), Ada (patrons de conceptions), C/C++, Lua (module Torch), Latex, bash, PowerShell, OCaml, Prolog				
Librairies	Torch, Theano (notions), Wapiti, MultiR, MaltParser				
Systèmes	Windows (7, 10) , Linux (Ubuntu, Mint, Debian, Arch)				
Outils	GIT, maven, JIRA, Eclipse, Microsoft Office, LibreOffice				
Méthodologie	Patrons de conceptions, programmation générique, Scrum				

Langues

	Comprendre		Parler		Écrire	
	Écouter	Lire	Converser	S'exprimer		
Français	C2	C2	C2	C2	C2	langue maternelle
Anglais	B2	C1	B2	B2	C1	

Projets libres open source

- SEM Segmenteur Étiqueteur Markovien. Librairie python pour l'annotation. Librairie créée durant mon stage de licence et enrichie régulièrement depuis. Permet l'annotation PoS, le chunking et en entités nommées. Export en plusieurs formats pour interopérabilité. [\[github\]](#)
- GAAL Generic Ada Automata Library. Librairie d'automates codée en Ada. Implémente notamment l'algorithme k-RI. Librairie créée durant mon stage de master. [\[github\]](#)

Responsables récents

- Gaël Lejeune Responsable pour l'[ATER à Sorbonne-Université](#). Maître de conférences. Contact : gael.lejeune@paris-sorbonne.fr
- Yannick Parmentier Responsable scientifique du partenaire LIFO pour le [post-doctorat du projet ANR PARSEME-FR](#). Maître de conférences. Contact : yannick.parmentier@loria.fr
- Éric Villemonte de la Clergerie Responsable scientifique pour le [post-doctorat du projet ANR PARSEME-FR](#) (Inria). Chargé de recherche. Contact : eric.de_la_clergerie@inria.fr

Publications

Articles de revue avec comité de lecture

DINARELLI, Marco et DUPONT, Yoann.

Modélisation de dépendances entre étiquettes dans les réseaux neuronaux récurrents.

In: Revue TAL. 2017, vol. 58, no 1. (*accepté*)

Communications dans des conférences internationales

MARTIN, Louis, MULLER, Benjamin, ORTIZ SUÁREZ, Pedro Javier, DUPONT, Yoann et al.

CamemBERT: a Tasty French Language Model.

In: LREC, 2020. (*accepté*)

ORTIZ SUÁREZ, Pedro Javier, DUPONT, Yoann, MULLER, Benjamin, et al.

Establishing a New State-of-the-Art for French Named Entity Recognition.

In: LREC, 2020. (*accepté*)

DUPONT, Yoann, DINARELLI, Marco and TELLIER, Isabelle.

Label-Dependencies Aware Recurrent Neural Networks ([Prix du meilleur programme CICling 2017](#)).

In: CICling. 2017.

[article](#)

TELLIER, Isabelle, MAKHLOUF, Zineb and DUPONT, Yoann.

Sequential Patterns of POS Labels Help to Characterize Language Acquisition.

In: DMNLP @ PKDD/ECML. 2014. p. 129-142.

[article](#)

TELLIER, Isabelle, DUPONT, Yoann, ESHKOL-TARAVELLA, Iris and WANG, Ilaine.

Adapt a Text-Oriented Chunker for Oral Data: How Much Manual Effort Is Necessary?

In: IDEAL. 2013. p. 226-233.

[article](#)

TELLIER, Isabelle and DUPONT, Yoann.

How Symbolic Learning Can Help Statistical Learning (and vice versa).

In: RANLP 2013. p. 649-658.

[article](#)

Communications dans des conférences nationales

DUPONT, Yoann.

Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique (Prix du meilleur article RECITAL 2017**).**

In: TALN-RECITAL. 2017.

[article](#)

DUPONT, Yoann, DINARELLI, Marco et TELLIER, Isabelle.

Réseaux neuronaux profonds pour l'étiquetage de séquences.

In: TALN-RECITAL. 2017.

[article](#)

MAKHLOUF, Zineb, DUPONT, Yoann, et TELLIER, Isabelle.

Caractériser l'acquisition d'une langue avec des patrons d'étiquettes morpho-syntaxiques.

In: JADT. 2014.

[article](#)

TELLIER, Isabelle, DUPONT, Yoann, ESHKOL-TARAVELLA, Iris et WANG, Ilaine.

Peut-on bien chunker avec de mauvaises étiquettes POS ?

In: TALN. 2014. p. 125-136.

[article](#)

TELLIER, Isabelle et DUPONT, Yoann.

Apprentissage symbolique et statistique pour le chunking : comparaison et combinaisons.

In: TALN-RECITAL 2013.

[article](#)

CONSTANT, Matthieu, TELLIER, Isabelle, DUCHIER, Denys, DUPONT, Yoann, Anthony Sigogne, Sylvie Billot
Intégrer des connaissances linguistiques dans un CRF: application à l'apprentissage d'un segmenteur-étiqueteur du français.

In: TALN 2011. p. 321.

[article](#)

Autres publications

DUPONT, Yoann

Un corpus libre, évolutif et versionné en entités nommées du Français.

In: TALN 2019 (poster).

[article](#)

DUPONT, Yoann, DINARELLI, Marco, TELLIER, Isabelle and LAUTIER, Christian.

Structured Named Entity Recognition by Cascading CRFs.

In: CICling. 2017 (poster).

[article](#)

DUPONT, Yoann et PLANCQ, Clément.

Un étiqueteur en ligne du Français.

session démonstration de TALN-RECITAL, 2017, p. 15 (démonstration).

[article](#)

DUPONT, Yoann, TELLIER, Isabelle, LAUTIER, Christian, et DINARELLI, Marco.

Extraction automatique d'affixes pour la reconnaissance d'entités nommées chimiques.

In: EGC. 2016 (poster).

[article](#)

DUPONT, Yoann et TELLIER, Isabelle.

Un reconnaiseur d'entités nommées du Français.

In TALN-RECITAL. 2014. p. 40-41. (démonstration)

[article](#)

TELLIER, Isabelle, DUPONT, Yoann et COURMET, Arnaud.

Un segmenteur-étiqueteur et un chunker pour le français (démonstration).

In: JEP-TALN-RECITAL 2012.

[article](#)