Pedro J. **Ortiz Suárez**

PH.D. STUDENT

pedro.ortiz@inria.fr | https://pjortiz.com/ | pjox

Interests

Language Modeling, Sequence Tagging, Name Entity Recognition, Parsing, Corpus Linguistics, Text Mining.

Education

Sorbonne Université — *Paris, France*

PH.D. IN COMPUTER SCIENCE — *Oct. 2018 - Present*

Ph.D. contract at Inria

Université de Paris 8, Vincennes–Saint-Denis — *Saint-Denis, France*

B.A.SC. MATHEMATICS AND COMPUTER SCIENCE APPLIED TO HUMAN AND SOCIAL SCIENCES — *Nov. 2017 - Sep. 2018*

Minor in Linguistics

Université d'Aix Marseille — *Marseille, France*

M.SC. IN MATHEMATICS — *Sep. 2016 - Jun. 2017*

Full scholarship given by the LabEx Archimède

Universidad Nacional de Colombia — *Medellín, Colombia*

B.SC. IN MATHEMATICS — *Jan. 2012 - Jun. 2016*

Admitted with honors, third place in the admission test

Experience

Inria - ALMAnaCH Team — *Paris, France*

PH.D. STUDENT — *Nov. 2018 - Present*

- Research in text mining and deep learning.

Sorbonne Université — *Paris, France*

GRADUATE TEACHING ASSISTANT — *Sep. 2019 - Présent*

- Discrete Mathematics, TD 40h.
- Eléments de programmation 2, TD 30h.

Inria - ALMAnaCH Team — *Paris, France*

RELAIS THÈSE — *Oct. 2018*

- Research in text mining and deep learning.

Inria - ALMAnaCH Team — *Paris, France*

INTERN — *Apr. 2018 - Sep. 2018*

- Research in text mining and deep learning.

Institut de Mathématiques de Marseille, I2M — *Marseille, France*

INTERN — *Apr. 2017 - Jun. 2017*

- Research internship in Complex Algebraic Geometry.

Universidad Nacional de Colombia — *Medellín, Colombia*

TEACHING ASSISTANT DISCRETE MATHEMATICS — *Aug. 2015 - Jun. 2016*

- Discrete Mathematics, TD 128h.

Publications

[1] Pedro Javier Ortiz Suárez, Benoît Sagot, Laurent Romary. *A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 2020, Association for Computational Linguistics, Online.

[2] Louis Martin*, Benjamin Muller*, Pedro Javier Ortiz Suárez*, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, Benoît Sagot. *CamemBERT: a Tasty French Language Model*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 2020, Association for Computational Linguistics, Online.

[3] Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, Abhishek Srivastava. *Building a User-Generated Content North-African Arabizi Treebank: Tackling Hell*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul 2020, Association for Computational Linguistics, Online.

[4] Louis Martin*, Benjamin Muller*, Pedro Javier Ortiz Suárez*, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, Benoît Sagot. *Les modèles de langue contextuels Camembert pour le Français : impact de la taille et de l'hétérogénéité des données d'entrainement*, 27e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2020), Jun 2020, Nancy, France.

[5] Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, Benoît Sagot. *Establishing a New State-of-the-Art for French Named Entity Recognition*, 12th International Conference on Language Resources and Evaluation (LREC 2020), May 2020, Marseille, France.

[6] Murielle Popa-Fabre, Pedro Javier Ortiz Suárez, Benoît Sagot, Éric de la Clergerie. *French Contextualized Word-Embeddings with a sip of CaBeRnet: a New French Balanced Reference Corpus*, 8th Workshop on the Challenges in the Management of Large Corpora (CMLC-8), May 2020, Marseille, France.

[7] Mohamed Khemakhem, Ioana Galleron, Geoffrey Williams, Laurent Romary, Pedro Javier Ortiz Suárez. *How OCR Performance can Impact on the Automatic Extraction of Dictionary Content Structures*, 19th annual Conference and Members' Meeting of the Text Encoding Initiative Consortium (TEI) -What is text, really? TEI and beyond, Sep 2019, Graz, Austria.

[8] Pedro Javier Ortiz Suárez, Benoît Sagot, Laurent Romary. *Asynchronous Pipelines for Processing Huge Corpora on Medium to Low Resource Infrastructures*, 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Jul 2019, Cardiff, United Kingdom.

*Equal contribution.

Talks

Séminaires du Lattice — *Paris, France*

Des Méthodes de TAL modernes pour l'Enrichissement de Documents — *22/09/2020*

École normale supérieure / Laboratoire Lattice

58th Annual Meeting of the Association for Computational Linguistics — *Online*

A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages — *06/07/2020*

Association for Computational Linguistics

7th Workshop on the Challenges in the Management of Large Corpora — *Cardiff, Wales, UK*

Asynchronous Pipelines for Processing Huge Corpora on Medium to Low Resource Infrastructures — *22/07/2019*

Cardiff University

10th International Conference on Historical Lexicography and Lexicology — *Leeuwarden, Netherlands*

Preparing the Dictionnaire Universel for Automatic Enrichment — *13/06/2019*

Fryske Akademy

GIG #3 : bring the cool back in the cloud — *Paris, France*

Reducing computation time by months by rewriting Bash scripts in Go — *24/03/2019*

Golang Paris

Honors & Awards

2016-2017	**Full master scholarship**, Granted by the LabEx Archimède, for academic excellence.	*Marseille, France*
2012	**Tuition payment exemption**, Top 15 of class, Granted by the Science Faculty Council, Universidad Nacional de Colombia, second academic period.	*Medellín, Colombia*
2012	**Tuition with honors**,Top 5 of class, granted by the Science Faculty Council, Universidad Nacional de Colombia, first academic period.	*Medellín, Colombia*

Languages

Spanish - Native (Colombia), **English** - Fluent (both written and spoken), **French** - Fluent (both written and spoken), **German** - Elementary (Limited speaking).

Skills and tools

• **Most experience:** C/C++, Python, Go, Matlab, Wolfram Mathematica, LaTeX, VS Code, TeXStudio, GNU/Linux, OS X, Git.
• **Moderate experience:** Raspberry Pi, JavaScript, PHP, HTML5, Java, Windows, SQL, MySQL, OpenSSH, GnuPG, Apache Server, CaddyServer, NumPy, GIMP, TensorFlow.