Pedro **Ortiz Suarez**

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 - Jun. 2021*

- Discrete Mathematics, TD 81h.
- 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] Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary and Benoît Sagot. *Ungoliant: An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus*, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9), Jul 2021, Leibniz-Institut für Deutsche Sprache, Online.

[2] Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, Mofetoluwa Adeyemi. *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets*, arXiv e-prints, Mar 2021, arXiv, Online.

[3] Pedro Javier Ortiz Suárez, Yoann Dupont, Gaël Lejeune, Tian Tian. *SinNer@Clef-Hipe2020 : Sinful adaptation of SotA models for Named Entity Recognition in French and German*, Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Sep 2020, CEUR-WS, Thessaloniki / Virtual, Greece.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.