Pedro **Ortiz Suarez**

PH.D. STUDENT

pedro@portizs.eu | portizs.eu | Pedro Ortiz Suarez | Pedro Ortiz Suarez | pjox | pjox | pjox |
@pjox13 | 0000-0003-0343-8852 | Pedro Ortiz Suarez | Pedro Ortiz Suarez | Pedro Ortiz Suarez

Education

Sorbonne Université - Inria - ALMAnaCH Team

Paris, France

PH.D. IN COMPUTER SCIENCE

Oct. 2018 - Apr. 2022

Topic: On Language Modeling and its Applications for Contemporary and Historical French
Advisers: Laurent Romary and Benoît Sagot
Funding: ANR BASNUM, public grant

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
- GPA: 16.5/20.0

Université d'Aix Marseille

Marseille, France

M.SC. IN MATHEMATICS

Sep. 2016 - Jun. 2017

Topic: Surfaces Canoniquement Plongées de Grands Degrés (Canonical Surfaces of High Degree)
Adviser: Xavier Roulleau

- Full scholarship granted by the LabEx Archimède
- GPA: 15.1/20.0

Universidad Nacional de Colombia

Medellín, Colombia

B.SC. IN MATHEMATICS

Jan. 2012 - Jun. 2016

Topic: A Brief Introduction to Arithmetic Geometry
Adviser: Juan Diego Vélez Caicedo

- Admitted with honors, third place in the admission test
- GPA: 4.2/5.0

Experience

Inria - ALMAnaCH Team

Paris, France

PH.D. STUDENT

Nov. 2018 - Apr. 2022

- Research in language modeling for contemporary and historical French
- Created and managed the OSCAR project
- Research in sequence tagging for contemporary and historical texts

Inria - ALMAnaCH Team

Paris, France

INTERN

Apr. 2018 - Sep. 2018

- Worked on a project with the Ministry of Work of France (Dares)
- Research in text mining and data extraction for enterprise contracts

Institut de Mathématiques de Marseille, I2M

Marseille, France

INTERN

Apr. 2017 - Jun. 2017

- Research internship in Complex Algebraic Geometry and Canonical Surfaces

Teaching

Sorbonne Université

Paris, France

GRADUATE TEACHING ASSISTANT

Sep. 2019 - Jun. 2021

- Discrete Mathematics, Tutorial 81h
- Eléments de programmation 2, Tutorials 30h
- Worked with interactive tools to facilitate teaching during the Covid- 19 pandemic

Universidad Nacional de Colombia

Medellín, Colombia

TEACHING ASSISTANT DISCRETE MATHEMATICS

Aug. 2015 - Jun. 2016

- Discrete Mathematics, Tutorials 128h
- Worked in a constraint resource environment with groups of more than 120 students

Research

Journal Publications

[1] **Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets.** Julia Kreutzer, Isaac Caswell, 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 Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, 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. *Transactions of the Association for Computational Linguistics*, volume 9, 2021 (to appear).

Conference Publications

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

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

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

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

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

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

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

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

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

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

*Equal contribution.

Open source software and corpora

- **OSCAR**: The Open Super-large Crawled Aggregated coRpus is a huge multilingual corpus obtained by language classification and filtering of the Common Crawl corpus using the Ungoliant architecture
- **CamemBERT**: A state-of-the-art language model for French based on the RoBERTa architecture and pretrained on the French subcorpus of the OSCAR corpus
- **FrELMo**: A language model for French based on the ELMo architecture and pretrained on the French subcorpus of the OSCAR corpus
- **Ungoliant**: A high-performance pipeline that provides tools to build a multilingual corpus from CommonCrawl. It is the current pipeline for the OSCAR corpus
- **Goclassy**: The original pipeline for buiding the OSCAR corpus, later replaced by Ungoliant

Reviewing

- ACL 2020, COLING 2020, EACL 2021, ACL 2021, EMNLP 2021, CHR 2021, ARR October 2021, ARR November 2021, JMDHD

Talks

23/11/2021	**Séminaires du Master Sciences du Langage,** Les Modèles de Langue pour le Français Contemporain et Historique	*Université Paris Nanterre, France*
29/06/2021	**Demo TALN,** CANTAL – Formats et ChAîNes de traitement de TAL	*Université de Lille, France*
22/09/2020	**Séminaires du Lattice,** Des Méthodes de TAL modernes pour l'Enrichissement de Documents	*ENS - Lattice, France*
06/07/2020	**58th Annual Meeting of the Association for Computational Linguistics,** A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages	*ACL 2020, Online*
22/07/2019	**7th Workshop on the Challenges in the Management of Large Corpora,** Asynchronous Pipelines for Processing Huge Corpora on Medium to Low Resource Infrastructures	*Cardiff University, UK*
13/06/2019	**10th International Conference on Historical Lexicography and Lexicology,** Preparing the Dictionnaire Universel for Automatic Enrichment	*Fryske Akademy, Netherlands*

24/03/2019 **GIG #3 : bring the cool back in the cloud,** Reducing computation time by months by rewriting Bash scripts in Go — *Golang Meeting Paris, France*

Miscellaneous

Project management

2019-now **OSCAR Project Coordination,** Head of the OSCAR project, supervision of a research engineer — *OSCAR Project*

Website and Infrastructure Administration

2019-now **OSCAR Project Website,** oscar-corpus.com — *ALMAnaCH team*
2018-now **Administration of the Inria ALmaNACH team Mattermost server,** traces1.inria.fr/mattermost/ — *ALMAnaCH Team*
2020 **CamemBERT Project Website,** camembert-model.fr — *ALMAnaCH team*

Skills

Machine learning NLP, PyTorch, Tensorflow, Tensorboard, Flair, spaCy, HF/Transformers, HF/Datasets (contributor), Fairseq
Programming Go, Rust, Python, Java, C/C++, PHP, JavaScript, Shell, Matlab, Wolfram Mathematica
Tools & DevOps LaTeX, Docker, Git, Apache Server, Mattermost, Grobid, MySQL, PostgreSQL, SQLLite, MongoDB, bash, Raspberry Pi
Languages **Spanish** (Native), **English** (Fluent), **French** (Fluent), **German** (Intermediary level)

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*