Pedro **Ortiz Suarez**

RESEARCHER

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

Research

Ph.D. Thesis

Sorbonne Université - Inria - ALMAnaCH Team *Paris, France*

PH.D. IN COMPUTER SCIENCE *Oct. 2018 - Jun. 2022*

Topic: A Data-driven Approach to Natural Language Processing for Contemporary and Historical French
Advisers: Laurent Romary and Benoît Sagot

Journal Publications

1. **Automatic extraction of materials and properties from superconductors scientific literature.** Luca Foppiano, Pedro Baptista Castro, Pedro Ortiz Suarez, Kensei Terashima, Yoshihiko Takano, Masashi Ishii *Science and Technology of Advanced Materials: Methods*, volume 3, 2023.
2. **Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets.** Julia Kreutzer, et al. (50+ authors) *Transactions of the Association for Computational Linguistics*, volume 10, 2022.

Conference Publications

1. **The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset.** Hugo Laurençon, et al. (50+ authors). *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, Nov 2022, New Orleans, United States.
2. **A Data-driven Approach to Named Entity Recognition for Early Modern French.** Pedro Ortiz Suarez and Simon Gabay. *The 29th International Conference On Computational Linguistics (COLING 2022)*, Oct 2022, Gyeongju, Republic of Korea.
3. **BERTrade: Using Contextual Embeddings to Parse Old French.** Loïc Grobol, Mathilde Regnault, Pedro Ortiz Suarez, Benoît Sagot, Laurent Romary and Benoit Crabbé. *13th International Conference on Language Resources and Evaluation (LREC 2022)*, May 2022, Marseille, France.
4. **From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French.** Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, Benoît Sagot. *13th International Conference on Language Resources and Evaluation (LREC 2022)*, May 2022, Marseille, France.
5. **Towards a Cleaner Document-Oriented Multilingual Crawled Corpus.** Julien Abadji, Pedro Ortiz Suarez, Laurent Romary and Benoît Sagot. *13th International Conference on Language Resources and Evaluation (LREC 2022)*, May 2022, Marseille, France.
6. **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.
7. **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.
8. **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.
9. **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.
10. **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.
11. **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.
12. **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.
13. **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.
14. **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.
15. **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.*

Pre-prints

1. **Semi-automatic staging area for high-quality structured data extraction from scientific literature.** Luca Foppiano, Tomoya Mato, Kensei Terashima, Pedro Ortiz Suarez, Taku Tou, Chikako Sakai, Wei-Sheng Wang, Toshiyuki Amagasa, Yoshihiko Takano, Masashi Ishii *CoRR, abs/2309.10923*,

Sep 2023.

2. **Perplexed by Quality: A Perplexity-based Method for Adult and Harmful Content Detection in Multilingual Heterogeneous Web Data.** Tim Jansen, Yangling Tong, Victoria Zevallos, Pedro Ortiz Suarez *CoRR, abs/2212.10440*, Dec 2022.
3. **BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.** BigScience Workshop. *CoRR, abs/2211.05100*, Nov 2022.
4. **Documenting Geographically and Contextually Diverse Data Sources: The BigScience Catalogue of Language Data and Resources.** Angelina McMillan-Major, Zaid Alyafeai, Stella Biderman, Kimbo Chen, Francesco De Toni, Gérard Dupont, Hady Elsahar, Chris Emezue, Alham Fikri Aji, Suzana Ilić, Nurulaqilla Khamis, Colin Leong, Maraim Masoud, Aitor Soroa, Pedro Ortiz Suarez, Zeerak Talat, Daniel van Strien, Yacine Jernite. *CoRR, abs/2201.10066*, Jan 2022.

Published 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 Experience

- ACL 2020, COLING 2020, EACL 2021, ACL 2021, EMNLP 2021, CHR 2021, ARR October 2021, ARR November 2021, JMDHD, ARR July 2022, EMNLP 2022, NEJLT, ACL 2023, EMNLP 2023.

Research Experience

DFKI GmbH - Speech and Language Technology Lab
Berlin, Germany

Researcher — *Jan. 2023 - Present*

- Worked for the OpenGPT-X project
- Continued the development of the OSCAR corpus

Universität Mannheim - Data and Web Science Group
Mannheim, Germany

Postdoctoral Researcher — *Mar. 2022 - Dec. 2023*

- Continued the development of the OSCAR corpus
- Production of new multilingual resources for NLP and web-based applications
- Teaching and project coaching

Inria - ALMAnaCH Team
Paris, France

Ph.D. Student — *Oct. 2018 - Feb. 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

Universität Mannheim
Mannheim, Germany

Lecturer — *Mar. 2022 - Present*

- Information Retrieval 2022, Lectures
- Introduction to Data science 2022, Lectures
- Web Mining 2022, Lectures
- Team Projects 2022, Coaching

Sorbonne Université
Paris, France

Graduate Teaching Assistant — *Sep. 2019 - Jun. 2021*

- Discrete Mathematics, Tutorials 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

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)

Education

Sorbonne Université - Inria - ALMAnaCH Team

Paris, France

PH.D. IN COMPUTER SCIENCE *Oct. 2018 - Jun. 2022*

Lab: Inria Paris' NLP, ALMAnaCH project-team
Doctoral School: Ecole Doctorale Informatique, Télécommunications et Electronique. EDITE de Paris (ED130)
Funding: ANR-18-CE38-0003 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 on the admission test
- GPA: 4.2/5.0

Talks

29/03/2022	**Data and Web Science Group Colloquium**, A Data-driven Approach to Natural Language Processing for Contemporary and Historical French	*Universität Mannheim, Germany*
03/12/2021	**Workshop en Philologie computationnelle: au delà de l'encodage du texte**, Reconnaître les entités nommées	*Université de Genève, Switzerland*
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*
2021-2022	**Big Science Area Chair and Founding Member**, Area chair of the Big Science Data Sourcing Working Group and founding member of the workshop	*Big Science*

WEBSITE AND INFRASTRUCTURE ADMINISTRATION

2019-now	**OSCAR Project Website**, oscar-corpus.org	*ALMAnaCH team*
2018-2022	**Administration of the Inria ALmaNACH team Mattermost server**, traces1.inria.fr/mattermost/	*ALMAnaCH Team*
2020	**CamemBERT Project Website**, camembert-model.fr	*ALMAnaCH team*

Volunteering

2020 **Participant in the Mission Covid-19 supporting the AP-HP**, Installation of Grobid and entity-fishing on the AP-HP (Paris hospitals) insfrastructure during the COVID-19 quarantine in France *Mission Covid-19*

Outreach

Oct. 2019 **Mentor at the RJMI at Inria Paris**, Young Women Maths/Computer Science Days *Inria 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*