2023-06448 - PhD Position F/M Fine-Grained Language Classification of Very Large Corpora

**Contract type:** Fixed-term contract

**Level of qualifications required:** Graduate degree or equivalent

**Fonction:** PhD Position

**Context**

This position is part of Inria’s DEFI project COLaF (Corpus and Tools for the Languages of France), which is a collaboration between the Inria ALMAnaCH (Inria Paris center) and MULTISPEECH (Inria Nancy–Grand Est center) project teams. The objective of the project is to develop and make available digital linguistic technologies for the French-speaking world (native and non-native, contemporary or not, in France and outside of France, etc.) and the languages of France (all languages spoken in France: regional Romance and non-Romance languages, Creoles, immigrant languages, etc.), by contributing to the creation of inclusive data corpora, models, and software bricks. ALMAnaCH focuses on text and MULTISPEECH on multimodal speech. The two main objectives of this project are:

- **The collection of massive and inclusive French-speaking data corpora:** The aim is to build very large textual and speech corpora, with rich metadata to improve the robustness of models in the face of linguistic variation, with a particular focus on geographic-dialectal variation in the context of the French-speaking world. Diachronic, diatopic and diastatic variations will be important elements to take into account in the corpus collection.

- **The development and provision of inclusive language technologies:** This includes but is not limited to the production of annotated resources (parts of speech, syntax, named entities) and the extraction and reformalization of structured data (e.g., dictionaries), with a view to training language models that promote the consideration of linguistic variety in France and in the **Francophonie**

The PhD will be co-supervised by Benoit Sagot (Project-team ALMAnaCH, Inria Paris, France) and Pedro Ortiz Suarez (Speech and Language Technology team, DFKI GmbH, Berlin, Germany), co-founders of the OSCAR project (Open Super-large Crawled Aggregated coRpus). It will be funded by ALMAnaCH, Inria Paris.

The PhD student will work at Inria, within the ALMAnaCH project-team, in collaboration with the whole COLaF team. Within the team, they will work with the starting researcher (SRP) who co-manages the project with Benoît Sagot and the corpus engineers in order to identify requirements for the COLaF dataset, in terms of metadata, language diversities and exploitability.

**Assignment**

**Subject**

Natural Language Processing (NLP) has undergone staggering changes over the past decade thanks to neural approaches and in particular in recent years. This is thanks to multiple factors including advances in computational resources making it possible to train large-scale models with billions of parameters and the creation of large-scale datasets on which to train models. Neural language models have provided significant boosts across NLP tasks (Kenton and Toutanova, 2019; Lewis et al., 2020; Chowdhery et al., 2022; Scao et al., 2022), and although the majority of research is carried out on English, an increasing attention is being paid to other languages, including those for which fewer resources are available (Haddow et al., 2022; Zhu et al., 2023).

Multilinguality in particular has been an important focus of research in low-resource languages, providing gains across tasks (Fan et al., 2021; Muller et al., 2021; Costa-jussá et al., 2022).

In this context, the ability to leverage massive and clean corpora has shown promises of performance gains (Ortiz Suárez, Romary, and Sagot, 2020). The OSCAR Project3 has been aiming to provide web-based multilingual resources (Ortiz Suárez, Sagot, and Romary, 2019) and datasets for Machine Learning (ML) and Artificial Intelligence (AI) applications (Ortiz Suárez, Romary, and Sagot, 2020). The project focuses specifically in providing large quantities of unannotated raw data that is commonly used in the pre-training of large deep learning models.

Specifically relevant to COLaF, the OSCAR project has developed high-performance data pipelines (Abadi et al., 2021) specifically conceived to classify and filter large amounts of web data while putting special attention in improving the data quality of web-based corpora (Abadi et al., 2022) as well as providing data for low-resource languages (Kreutzer et al., 2022), so that these new ML/AI technologies are accessible to as many communities as possible. One of the on-going areas of research within the OSCAR project is that of language classification, this is mostly due to the fact that, even though recent research has proven that it is possible to develop language identification models capable of distinguishing between hundreds (NLLB Team et al., 2022) or even thousands of languages (Siddhant et al., 2022), this approaches have also been shown to have serious limitations, be it in
terms of the architecture itself (Caswell et al., 2020) or on the data used to train the model (Burchell et al., 2023). These findings suggest that, simpler non-neural approaches might also be worth exploring (Jauhiainen, Jauhiainen, and Lindén, 2022). As a consequence of the performance limitations of language classification, data can be messy and lead to unexpected outcome: recent research (Briakou, Cherry, and Foster, 2023) has shown that the presence of undetected bilingual example and—more importantly—translation pairs, in the training data of Large Language Model (LLM) such as PaLM is partially responsible for the model’s translation capabilities.

As such, fine linguistic classification (LC) remains an important challenge to solve. Outside of language classification, work on code-switching identification, multilingual documents detection (Caswell et al., 2021), genre classification, hate-speech detection (Jansen et al., 2022) across languages and variation such as diatopic, diastratic or diachronic ones, has yet to demonstrate an ability to work at scale on billions of tokens despite leveraging lower resource. Within the current pipeline, OSCAR has to leave aside around 12 TB of data (early 2019 estimate), in which we find a mix of voluntarily rejected data as well as unclassifiable text.

This PhD will explore the question of fine-grained linguistic classification in the context of both inference efficiency (PBs of data have to be treated) and accuracy. There are 3 main goals:

- the ability to more accurately detect and sort low-resource languages at scale, with a particular focus on France’s regional languages such as Occitan, Picard and Creole languages spoken in France,
- the ability to detect variation across speakers, based on geographical, chronological or sociological criteria, in order to help model training focus on the diversity of speakers within a single language,
- the ability to extract other markers such as code-switching or bilingual data.

To achieve this, a number of different approaches could be explored, from partially class-free classification, unsupervised and supervised classification in the context of imbalanced dataset, hierarchical clustering and classification, and more. Both statistical and neural approaches could be combined, sometimes even with basic pattern-matching approaches.

Main activities
The main activities of the PhD will include:

- keeping up-to-date with related work on the topic
- carrying out research on the topic outlined above, both in the development of new ideas, positioning with respect to related work and validation of the methodology via experiments and analysis
- contribute to the identification requirements for the COLaF dataset, in terms of metadata, language diversities and exploitability
- manage the interactions with OSCAR users, as well as the regular creation of new versions of OSCAR and their publication
- the presentation of work both internally to colleagues and externally in the form of conference/journal/workshop papers and in the final PhD thesis
- interacting and exchanging with colleagues on NLP topics

Skills
Candidates should have a good level in programming (python required, RUST would be a plus as plus), experience with neural networks and an interest in natural language processing. A good written and spoken level of English is required.

Benefits package
- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Teleworking
- Flexible organization of working hours (after 12 months of employment)
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Remuneration
The gross salary will be 2051€/month for the first two years and then 2158€/month for the third year.
Contacts

- **Inria Team**: [ALMANACH](#)
- **PhD Supervisor**: Sagot Benoit / Benoit.Sagot@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

The keys to success

The position is a 3-year funded PhD position of starting on the 1st October 2023 (or possibly on the 1st November). Candidates should have a Master 2 or equivalent (e.g. engineering school) in computer science (speciality artificial intelligence, machine learning or natural language processing). They should have a good level in programming (python required, RUST would be a plus as plus), experience with neural networks and an interest in natural language processing. A good written and spoken level of English is required.

We are looking for highly motivated candidates with a strong background in NLP, machine learning and an interest in linguistics and language, and in particular in language diversity and languages of France. Ideally, candidates should be able to show initiative, creativity and have a good eye for analysis of data and results.

Warning: you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

In your application (which can be in English or in French), please include:
- CV
- Letter of motivation
- Letters of recommendation
- Optionally an example of your previous written work (if possible related to NLP), for example a master's thesis, research paper, etc.
- Master's notes

Defence Security :

This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST). Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.