La Rochelle Université
Institut Littoral Urbain
Durable Intelligent

PhD position description

Title of the thesis project
Robust-to-noise information extraction, unifying challenges of optical character recognition (OCR) and automatic speech recognition (ASR)

La Rochelle University Research Unit: L3i

Name of the LRUniv supervisor
Mickaël Coustaty / Cyrille Suire

Main research field
Computer Science

Secondary research field(s):
Natural Language Processing, Information Extraction, Deep Learning

Keywords *(6 max)*
Robust NLP, OCR, ASR, information extraction

Scientific description of the research project

Scientific context:
The growing digitization of written and oral content has made Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR) essential in cultural heritage preservation, media accessibility, legal documentation, knowledge management and information retrieval. However, the outputs generated by these systems are inherently noisy: OCR is affected by document degradation, layout complexity or poor scanning quality, while ASR suffers from background noise, overlapping speech or non-standard oral expressions. Despite significant progress, it remains pervasive, and imperfections directly impact natural language downstream tasks where data quality is a key prerequisite. Although OCR and ASR face many similar error phenomena, their correction has mostly been studied in isolation, resulting in a lack of unified methodologies.

Objectives:
- Compare and analyse existing post-correction methods in OCR and ASR and potential for cross-domain adaptation.
- Develop unified approaches for post-correction that leverage the shared error patterns between OCR and ASR.
- Enable robust information extraction from noisy OCR and ASR outputs by designing strategies that mitigate the propagation of recognition errors into downstream NLP tasks.

Scientific challenges:
- **Heterogeneity of noise sources:** OCR errors are generated from visual artifacts while ASR errors are acoustic, a unified framework must generalize across modalities.
- **Domain adaptation:** OCR/ASR models often struggle on domain-specific datasets (e.g., historical texts, administrative documents, technical reports, scientific papers…) requiring correction methods that adapt to varying contexts.
- **Complex error structures**: beyond character and subword substitution, OCR/ASR introduce higher-level disruptions (mis-segmentation, overlapping text blocs or speech, layout misinterpretation) that complicate correction.

- **Evaluation difficulties:** classical metrics such as Character Error Rate (CER) or Word Error Rate (WER) fail to fully capture the impact of errors on downstream information extraction, that necessitate new evaluation methods.
- **Scalability:** correction methods must be applicable to large-scale corpora and adaptable to new data without full retraining.

To tackle these challenges, the thesis will explore a combination of:

- **Comparative state-of-the-art analysis**: systematic benchmarking of existing OCR and ASR post-correction methods on heterogeneous corpora.
- **Unified modeling approaches**: leveraging neural architectures (e.g., sequence-to-sequence models, transformers, multilingual pre-trained LLMs) that can learn correction patterns across both modalities.
- **Hybrid methods**: integrating symbolic rules, edit distance algorithms, and domain-specific lexicons with machine learning models to improve robustness.
- **Error modeling and simulation**: designing artificial noise injection techniques to train models on synthetic but realistic OCR/ASR-like errors, thus improving generalization.
- **Evaluation frameworks**: extending standard CER/WER with task-oriented metrics reflecting the quality of downstream information extraction and retrieval.

This thesis helps to overcome the current limitations of automatic correction of texts produced by OCR and ASR systems by proposing a unified approach, which represents a significant scientific advance. In fact, in-depth analysis of the similarities and differences between OCR and ASR errors will provide a better understanding of how these two fields can intersect. This project will enable the development of more robust methods based on multidisciplinary knowledge from natural language processing, signal processing, and image processing. The expected results will thus offer new perspectives in the development and use of multimodal language models, contributing to the evolution of generative AI in both language processing and signal processing. With the rise of multimodal databases (text, image, audio, video), this thesis could inspire the creation of tools capable of simultaneously exploiting data from various sources to extract more relevant information. The thesis is expected to deliver a contribution to the bridging of OCR and ASR research communities and opening new research avenues in multimodal NLP.

PhD student profile and skills required

The candidate should hold a master's degree in computer science or a related field. She/he should have a strong background in NLP with an interest in text processing and multimodal data (text, speech, document images). Familiarity with generative AI methods (e.g., large language models, text-to-text generation, deep learning, fine-tuning strategies) will form a strong asset.

Societal and economic challenges and contributions

The thesis project has significant socio-economic repercussions, such as cost reductions in industries in several sectors, including media, document management, and history. More reliable systems will reduce the need for human intervention in the manual correction of transcription errors. The reliability of OCR and ASR systems also improves access to information in documents and audio transcriptions, particularly those that are of poor quality or damaged. This improvement in accessibility would also have a positive impact on people with hearing or visual impairments, facilitating their access to audio content via reliable written and oral transcriptions. Improvements in these technologies would therefore make information more accessible to marginalized populations by strengthening their digital inclusion. The impact of this project will also affect education and research and will help researchers, teachers, and students with more efficient access to data. More reliable automatic correction of digitized documents and oral transcriptions will facilitate the automatic translation of digital content and remote collaborative work, respectively. This thesis also contributes to the efficiency of public administration and justice by improving the digitization of administrative documents and the transcription of minutes, while ensuring greater transparency and traceability.

OCR and ASR corrections help to preserve and disseminate cultural heritage. On the one hand, more reliable OCR technologies make it possible to digitize old documents and heritage archives that are often fragile and vulnerable to physical damage, thus facilitating their dissemination to a wider audience, including researchers, historians, students, and culture enthusiasts, regardless of their geographical

location. Putting these heritage collections online requires the extraction and analysis of their texts, facilitating their indexing, translation, and accessibility. In addition, this project will enable more accurate transcription of historical voice recordings, which are often degraded by background noise, interruptions, or poor sound clarity. This will improve accessibility to historical narratives, public speeches, period testimonies, and oral traditions that would otherwise be difficult to exploit.

Commitments in terms of co-supervision and work environment of the thesis *(human, material and, financial contributions that will be provided by the laboratories)*

- **Human support**: the doctoral student will benefit from solid human support, mainly provided by the thesis supervisors and co-supervisor. Regular meetings will be organized to monitor the progress of the project, resolve any difficulties, and guide the doctoral student in their research. These exchanges will also promote the sharing of experiences, the transmission of knowledge, and the development of the doctoral student's skills. In addition, collaboration with other researchers within the Image and Content team and the two other teams at the L3i laboratory, namely E-Adapt and Model and Knowledge, will be encouraged to promote a multidisciplinary approach and enrich the doctoral student's experience.
  In the University of Ljubljana, the student will be a member of Machine Learning and Language Technologies Lab and will get support of the other PhD students, researchers, and other staff. In addition, the student will be encouraged to participate in activities (trainings, collaborations, and research) of Centre of Excellence in Artificial Intelligence for Digital Humanities.
- **Material support:** L3I undertakes to provide doctoral students with the necessary material resources, such as computer equipment and other appropriate equipment. A budget is also allocated to enable doctoral students to participate in conferences and scientific events, thereby promoting their professional development and integration into the academic community. In addition, doctoral students may receive additional funding through existing projects led by their thesis supervisor and related to the topic of their thesis. Doctoral students will also have access to a complete range of material infrastructure to carry out their work. This includes access to high-performance computing servers equipped with the latest technologies for AI work within the L3i laboratory or at the regional level (. Dedicated workspaces, such as meeting rooms, will also be available to foster an environment conducive to research. Access to databases, specialized software, and essential documentary resources will be provided to enable in-depth analysis and the production of high-quality results.
  In the University of Ljubljana, the student will have available all material resources of other PhD students, including access to university's considerable HPC computing facilities, access to national supercomputing grid and novel AI factory supercomputer of Slovenia. A student will also have access to other resources available to the university's staff such as meeting rooms, software, parking space, recreation facilities, etc.

Project follow-up *(describe the steps that will be taken to ensure the project's long-term viability)*

To ensure the long-term viability of the project, a comprehensive follow-up strategy will be implemented. This includes establishing a clear communication framework between all parties, with regular virtual meetings and progress reports to maintain alignment across the group. A shared digital workspace will be used for collaborative documentation, milestone tracking, and resource sharing (some of the tools were mentioned earlier). Weakly meetings will be set online with both supervisors using MS Teams, where the student is located at the time. Additionally, contingency plans will be developed to address potential challenge. Periodic evaluations will be conducted to assess academic progress, research impact, and international collaboration effectiveness, ensuring the project remains on track and sustainable throughout the PhD duration and beyond. This of course includes the "CSI" yearly assessment in France. The long experience of successful collaboration between the two supervisors will certainly support the smooth execution of the PhD project.

Financing envisaged for the continuation of the project

As mentioned previously, a clear plan is to further develop this work into HEU proposals. A joint proposal is currently under preparation for topic HORIZON-CL2-2025-01-HERITAGE-03 with deadline in September 2025, in the context of the European Collaborative Cloud for Cultural Heritage (ECCCH), in a

project focused on the analysis of digitised documents of the Ottoman (very prone to OCR errors). The context of the European Centre of Excellence in AI4DH is very in terms of upcoming project proposals as it is one of its inherent goals. The connection and early connection with this PhD proposal would embark LRUniv further in this context. A proposal is planned in 2026 for an MSCA Marie Curie doctoral network, as well as twinning project (Widening). These would offer natural venue to pursue and largely expand the collaboration.
Applications for further projects in the area of digital humanities, relevant for the planned research topic, are in progress, such as EU ERC application for research in digital folkloristics, and application for EU HE infrastructure in electronic lexicography,