La Rochelle Université
Institut Littoral Urbain Durable Intelligent

HR EXCELLENCE IN RESEARCH

PROPOSITION DE SUJET POUR UN CONTRAT DOCTORAL/
Clôture des candidatures le : ……..

Laboratoire : L3i

Titre de la thèse :

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

Direction de la thèse *directeur·trice·s (grade, HDR) et éventuels co-directeur·trice·s*

Suire Cyrille, MCF, 50%

Mickael Coustaty, MCF, HDR, 50%

Adéquation scientifique avec les priorités de l'établissement

Cette thèse propose des solutions pour extraire, structurer et agréger des données issues de sources hétérogènes, telles que des livres, des archives audio ou vidéo, des documents patrimoniaux, etc. Elle vise ainsi à contribuer principalement à la préservation et à la valorisation du patrimoine numérique. Cependant, ce projet dépasse largement les seules problématiques patrimoniales et s'inscrit pleinement dans les objectifs de durabilité urbaine côtière intelligente (LUDI).

Tout d'abord, ce travail a pour ambition d'élargir l'accès à une grande variété d'informations pertinentes et précieuses. Grâce à la conservation et à la promotion de documents écrits et oraux, le projet garantit la transmission et la préservation de connaissances essentielles, s'inscrivant ainsi dans une démarche de durabilité. Par ailleurs, il introduit des solutions fondées sur l'intelligence artificielle, améliorant ainsi la gestion intelligente des données textuelles dans les systèmes numériques. Les technologies de correction post-OCR et post-reconnaissance vocale (ASR) contribueront à la préservation des documents anciens et des archives vocales numérisées, souvent sujets à une dégradation physique. La correction automatique permet de créer des copies numériques plus précises, assurant la pérennité des archives patrimoniales et historiques, tout en réduisant le besoin de traitements manuels intensifs, ce qui diminue la charge de calcul et l'empreinte énergétique des opérations de numérisation et de transcription.

Ce travail propose également des solutions utiles pour les activités de recherche et d'innovation dans des domaines d'application pertinents pour l'institution. L'amélioration de l'accessibilité des informations contenues dans les archives municipales et les registres publics, grâce à des corrections automatiques fiables, facilite l'accès des citoyens à des informations précises et transparentes. Les villes modernes s'appuient de plus en plus sur une gestion efficace de grandes quantités de documents administratifs et juridiques. Les technologies post-OCR et post-ASR améliorent l'accessibilité des documents, favorisant ainsi le développement de services essentiels

pour les villes de demain, tels que les procédures administratives, la gestion des registres publics ou encore les enquêtes législatives, judiciaires ou environnementales.

Enfin, l'intersection entre les sources textuelles natives (articles de presse, livres) et les sources audiovisuelles présente un intérêt dans de nombreux secteurs d'activité. Par exemple, nous prévoyons de développer ces approches pour des problématiques liées au littoral et à l'environnement. Ce travail pourrait contribuer à l'analyse des discours associés à des controverses environnementales (incidents de pollution côtière, projets d'aménagement ou de construction controversés, gestion des ressources naturelles, etc.), en offrant un accès unifié et fiable à des sources documentaires hétérogènes par leur nature et leur format, avec un processus de création maîtrisé. Les questions de recherche identifiées et sur lesquelles ce travail a un impact concernent, par exemple, la méthodologie de consolidation des corpus d'analyse médiatique ou l'analyse de la propagation des arguments dans le débat public (lobbying, fake news, etc.).

Descriptif du sujet *(enjeux scientifiques, applicatifs, sociétaux…)*

La numérisation croissante des contenus écrits et oraux a rendu la **reconnaissance optique de caractères (OCR)** et la **reconnaissance automatique de la parole (ASR)** essentielles dans des domaines tels que la préservation du patrimoine culturel, l'accessibilité des médias, la documentation juridique, la gestion des connaissances et la recherche d'informations. Cependant, les résultats générés par ces systèmes sont intrinsèquement bruités : l'OCR est affectée par la dégradation des documents, la complexité des mises en page ou la mauvaise qualité de numérisation, tandis que l'ASR souffre du bruit de fond, des chevauchements de parole ou des expressions orales non standard. Malgré les progrès significatifs réalisés, ces imperfections persistent et impactent directement les tâches aval de traitement automatique du langage naturel, où la qualité des données est un prérequis clé. Bien que l'OCR et l'ASR partagent de nombreux phénomènes d'erreur similaires, leur correction a principalement été étudiée de manière isolée, ce qui a conduit à un manque de méthodologies unifiées.

Objectifs

- Comparer et analyser les méthodes existantes de post-correction en OCR et ASR, ainsi que leur potentiel d'adaptation inter-domaines.
- Développer des approches unifiées de post-correction exploitant les schémas d'erreur partagés entre l'OCR et l'ASR.
- Permettre une extraction robuste d'informations à partir de données bruitées issues de l'OCR et de l'ASR, en concevant des stratégies limitant la propagation des erreurs de reconnaissance dans les tâches aval de TALN.

Défis scientifiques

- **Hétérogénéité des sources de bruit** : Les erreurs d'OCR proviennent d'artéfacts visuels, tandis que celles de l'ASR sont acoustiques. Un cadre unifié doit généraliser ces modalités.
- **Adaptation aux domaines** : Les modèles OCR/ASR peinent souvent sur des jeux de données spécifiques (textes historiques, documents administratifs, rapports techniques, articles scientifiques, etc.), nécessitant des méthodes de correction adaptables à des contextes variés.
- **Structures d'erreur complexes** : Au-delà des substitutions de caractères ou de sous-mots, l'OCR et l'ASR introduisent des perturbations de haut niveau (mauvaise segmentation, chevauchement de blocs de texte ou de parole, interprétation erronée de la mise en page), ce qui complique la correction.

- **Difficultés d'évaluation** : Les métriques classiques comme le **taux d'erreur par caractère (CER)** ou le **taux d'erreur par mot (WER)** ne capturent pas pleinement l'impact des erreurs sur l'extraction d'informations en aval, ce qui nécessite de nouvelles méthodes d'évaluation.
- **Passage à l'échelle** : Les méthodes de correction doivent être applicables à des corpus à grande échelle et adaptables à de nouvelles données sans réentraînement complet.

Approches proposées

Pour relever ces défis, la thèse explorera une combinaison de méthodes :

- **Analyse comparative de l'état de l'art** : Benchmark systématique des méthodes existantes de post-correction OCR et ASR sur des corpus hétérogènes.
- **Approches de modélisation unifiée** : Utilisation d'architectures neuronales (modèles séquence-à-séquence, transformers, LLMs multilingues pré-entraînés) capables d'apprendre des schémas de correction transmodaux.
- **Méthodes hybrides** : Intégration de règles symboliques, d'algorithmes de distance d'édition et de lexiques spécifiques au domaine avec des modèles d'apprentissage automatique pour améliorer la robustesse.
- **Modélisation et simulation des erreurs** : Conception de techniques d'injection artificielle de bruit pour entraîner les modèles sur des erreurs synthétiques mais réalistes, améliorant ainsi leur généralisation.
- **Cadre d'évaluation** : Extension des métriques CER/WER avec des indicateurs orientés tâche, reflétant la qualité de l'extraction et de la récupération d'informations en aval.

Contributions attendues

Cette thèse vise à surmonter les limitations actuelles de la correction automatique des textes produits par les systèmes OCR et ASR en proposant une approche unifiée, ce qui représente une avancée scientifique significative. En effet, une analyse approfondie des similitudes et des différences entre les erreurs d'OCR et d'ASR permettra de mieux comprendre comment ces deux domaines peuvent s'enrichir mutuellement. Ce projet permettra le développement de méthodes plus robustes, fondées sur des connaissances multidisciplinaires en traitement automatique du langage, traitement du signal et traitement d'image.

Les résultats attendus offriront de nouvelles perspectives dans le développement et l'utilisation de modèles de langage multimodaux, contribuant ainsi à l'évolution de l'IA générative, tant dans le traitement du langage que dans celui du signal. Avec l'essor des bases de données multimodales (texte, image, audio, vidéo), cette thèse pourrait inspirer la création d'outils capables d'exploiter simultanément des données issues de diverses sources pour en extraire des informations plus pertinentes. Elle devrait également contribuer à rapprocher les communautés de recherche en OCR et ASR, ouvrant de nouvelles voies de recherche en TALN multimodal.

Contexte partenarial *(cotutelle internationale, EU-CONEXUS, partenariat avec un autre laboratoire, une entreprise…)*

Ce projet s'inscrit dans le cadre d'une collaboration renforcée en **traitement automatique du langage naturel** entre **La Rochelle Université** (laboratoire L3i) et **l'Université de Ljubljana**. Les équipes dirigées par Antoine Doucet et Marko Robnik-Šikonja ont d'abord collaboré dans le cadre du projet H2020 Embeddia (2,99 M€, projet n°825153, 2019-2022) sur les plongements multilingues pour les langues moins représentées, où elles ont respectivement travaillé sur l'extraction d'informations et le développement de modèles de langage adaptés. Avec l'avènement des grands modèles de langage basés sur les transformers, cette proposition vise à renforcer cette collaboration dans un contexte totalement différent, mais confronté à une

problématique clé commune : la limitation des données. Ici, il ne s'agit plus de données en quantité limitée pour des langues peu dotées, mais de données affectées par des erreurs de transcription (qu'elles proviennent d'images, de textes ou de la parole). Dans les deux cas, les questions d'adaptation, d'affinage et de distillation de modèles restent pertinentes.

Cette collaboration se poursuit désormais dans le cadre du **centre européen d'excellence en intelligence artificielle pour les humanités numériques (AI4DH)**, piloté par Antoine Doucet et coordonné par Marko Robnik-Šikonja. Dans ce contexte, ce financement de thèse représentera le **premier cofinancement effectif** impliquant les deux universités, permettant d'élargir le groupe de chercheurs impliqués à La Rochelle.

Cette dynamique pourrait favoriser de futures propositions dans le cadre d'Horizon Europe, les deux équipes étant très engagées dans des projets collaboratifs, dans un domaine particulièrement actif, avec de nombreuses applications industrielles et sociétales émergentes dans le contexte de l'essor des modèles génératifs d'IA et des grands modèles de langage.

Impacts *(scientifiques, technologiques, socio-économiques, environnementaux, sociétaux…)*

Ce sujet de thèse présente des **répercussions socio-économiques significatives**, notamment en termes de réduction des coûts pour plusieurs secteurs industriels, tels que les médias, la gestion documentaire et l'histoire. Des systèmes plus fiables diminueront le besoin d'intervention humaine pour la correction manuelle des erreurs de transcription. La fiabilité accrue des systèmes OCR et ASR améliorera également l'accès à l'information contenue dans les documents et les transcriptions audio, en particulier ceux de mauvaise qualité ou endommagés. Cette amélioration de l'accessibilité aura un impact positif sur les personnes atteintes de déficiences auditives ou visuelles, en facilitant leur accès aux contenus audio grâce à des transcriptions écrites et orales plus fiables. Ainsi, ces avancées technologiques rendront l'information plus accessible aux populations marginalisées, renforçant leur inclusion numérique.

L'impact de ce projet s'étendra également aux **domaines de l'éducation et de la recherche**, en offrant aux chercheurs, enseignants et étudiants un accès plus efficace aux données. Une correction automatique plus fiable des documents numérisés et des transcriptions orales facilitera respectivement la traduction automatique des contenus numériques et le travail collaboratif à distance. Cette thèse contribuera aussi à l'efficacité de l'administration publique et de la justice, en améliorant la numérisation des documents administratifs et la transcription des procès-verbaux, tout en garantissant une meilleure transparence et traçabilité.

Les corrections OCR et ASR jouent également un rôle clé dans la **préservation et la diffusion du patrimoine culturel**. D'une part, des technologies OCR plus fiables permettront de numériser des documents anciens et des archives patrimoniales, souvent fragiles et vulnérables aux dégradations physiques, facilitant ainsi leur diffusion auprès d'un public plus large : chercheurs, historiens, étudiants et passionnés de culture, quelle que soit leur localisation géographique. La mise en ligne de ces collections patrimoniales nécessite l'extraction et l'analyse de leurs textes, ce qui en facilite l'indexation, la traduction et l'accessibilité. D'autre part, ce projet permettra une transcription plus précise des enregistrements vocaux historiques, souvent dégradés par des bruits de fond, des interruptions ou une mauvaise clarté sonore. Cela améliorera l'accès aux récits historiques, discours publics, témoignages d'époque et traditions orales, qui seraient autrement difficiles à exploiter.

Programme de travail du doctorant *(tâches confiées au doctorant)*

Le projet de thèse est organisé en **trois phases progressives**, chacune combinant développement méthodologique, validation expérimentale et diffusion scientifique.

Année 1 – Analyse et Fondations

La première année est consacrée à l'établissement des fondements scientifiques et expérimentaux.

Le ou la doctorant(e) devra :

- Réaliser une revue exhaustive de l'état de l'art sur la post-correction OCR/ASR et leur impact sur les tâches aval de TALN ;
- Effectuer une analyse fine des types d'erreurs selon les modalités (texte, image, parole) ;
- Préparer le protocole expérimental et constituer des corpus hétérogènes ;
- Implémenter des systèmes de référence et des simulateurs de bruit réalistes.

Cette phase garantit une compréhension solide des méthodes existantes et fournit le cadre comparatif nécessaire pour les contributions ultérieures.

Année 2 – Exploration et Contributions Cœur

La deuxième année est dédiée aux avancées méthodologiques.

Le ou la doctorant(e) devra :

- Explorer des approches hybrides de correction, combinant des composants symboliques et neuronaux ;
- Développer et évaluer des modèles unifiés de correction, capables de traiter le bruit indépendamment de la modalité d'acquisition ;
- Proposer et formaliser de nouvelles métriques pour évaluer la robustesse des tâches aval ;
- Effectuer un séjour de recherche international (académique et/ou industriel) pour renforcer les dimensions scientifique et appliquée du travail.

Cette phase constitue l'apport scientifique central de la thèse.

Année 3 – Consolidation et Diffusion

La troisième année est axée sur la consolidation, la généralisation et la valorisation.

Le ou la doctorant(e) devra :

- Étendre et affiner l'évaluation sur des domaines variés, des langues et des conditions de bruit différentes ;
- Adapter les modèles proposés à des environnements réels et aux jeux de données des partenaires ;
- Préparer des publications scientifiques, des versions logicielles et le manuscrit de thèse.

Cette phase vise à finaliser les résultats, garantir leur reproductibilité et maximiser leur impact.

Accompagnement du doctorant / Fonctionnement de la thèse *(accompagnement humain, matériel, financier, en particulier pour la prise en charge du fonctionnement de la thèse et des dépenses associées)*

Soutien humain

Le ou la doctorant(e) bénéficiera d'un **soutien humain solide**, principalement assuré par les **directeurs et co-directeurs de thèse**. Des **réunions régulières** seront organisées pour suivre l'avancement du projet, résoudre les éventuelles difficultés et guider le ou la doctorant(e) dans

ses recherches. Ces échanges favoriseront également le **partage d'expériences**, la **transmission de savoirs** et le **développement des compétences** du ou de la doctorant(e).

En outre, la **collaboration avec d'autres chercheurs** au sein de l'équipe **Image et Contenu** du laboratoire L3i, ainsi qu'avec les deux autres équipes, **E-Adapt** et **Modèle et Connaissance**, sera encouragée pour promouvoir une **approche multidisciplinaire** et enrichir l'expérience du ou de la doctorant(e).

À l'**Université de Ljubljana**, le ou la doctorant(e) intégrera le **Machine Learning and Language Technologies Lab** et bénéficiera du soutien des **autres doctorants, chercheurs et membres du personnel**. Il ou elle sera également encouragé(e) à participer aux **activités** (formations, collaborations, recherches) du **Centre d'excellence en intelligence artificielle pour les humanités numériques (AI4DH)**.

Soutien matériel

Le **laboratoire L3i** s'engage à fournir au ou à la doctorant(e) les **ressources matérielles nécessaires**, telles que du **matériel informatique** et d'autres équipements adaptés. Un **budget** est également alloué pour permettre la participation à des **conférences et événements scientifiques**, favorisant ainsi le **développement professionnel** et l'**intégration dans la communauté académique**. De plus, le ou la doctorant(e) pourra bénéficier de **financements complémentaires** grâce aux projets existants menés par son directeur de thèse et en lien avec le sujet de sa thèse.

Le ou la doctorant(e) aura également accès à une **infrastructure matérielle complète** pour mener à bien ses travaux, incluant :

- L'accès à des **serveurs de calcul haute performance** équipés des dernières technologies pour le travail en IA, au sein du L3i ou au niveau régional ;
- Des **espaces de travail dédiés**, comme des salles de réunion, pour favoriser un environnement propice à la recherche ;
- L'accès à des **bases de données**, des **logiciels spécialisés** et des **ressources documentaires essentielles** pour permettre des analyses approfondies et la production de résultats de qualité.

À l'**Université de Ljubljana**, le ou la doctorant(e) disposera de **toutes les ressources matérielles** mises à disposition des autres doctorants, notamment :

- L'accès aux **infrastructures de calcul haute performance (HPC)** de l'université ;
- L'accès au **réseau national de supercalculateurs** et au **nouveau supercalculateur AI Factory** de Slovénie ;
- L'utilisation d'autres ressources disponibles pour le personnel universitaire, telles que des **salles de réunion, des logiciels, des places de parking, des installations de loisirs**, etc.