UNIVERSITA DI CORSICA - PASQUALE PAOLI

ECOLE DOCTORALE "Environnement et Société"

BP 52, 20250 Corti – Tel : 04 95 37 23 22 / 45 02 20 – E.mail : ecole.doctorale@univ-corse.fr
Site : http://ed.univ-corse.fr

FICHE PROJET DE THESE pour ANNEE 2024-2025

Discipline du Doctorat	4200018 – Informatique
Mention du Doctorat	Informatique – 27°
Domaine scientifique principal	Informatique
Domaines scientifiques secondaires	Traitement Automatique des Langues
Entités de rattachement	UMR SPE / SISU
Direction de la thèse	Dr-HDR Paul-Antoine Bisgambiglia bisgambiglia_pa@univ-corse.fr Dr-HDR Estelle Retali-Medori medori_e@univ-corse.fr
Collaborations extérieures éventuelles envisagées	Universités de Strasbourg, Paris 8, Toulouse, Poitiers
Type de financement visé	Contrat doctoral
Connaissances et compétences requises chez l'étudiant	Master Informatique et/ou Linguistique Compétences : Linguistique, TAL, programmation, machine Learning
Titre de la thèse	**Adaptation des modèles de type LLM et conception d'outils pour le Traitement Automatique des Langues appliqué au corse**
Abstract 1	L'intérêt de ce sujet de thèse est d'adapter les outils actuels du TALN comme les « Large Language Model » pour saisir le contexte et le sens des mots, afin de créer des outils numériques au service de la revitalisation du corse et ainsi compléter les outils numériques de la BDLC.
Abstract 2	L'objectif est de simplifier l'accès aux données de la BDLC en permettant son interrogation en langage naturel. Les récentes avancées en intelligence artificielle, dont bénéficient les modèles de langues de type OpenAI/ChatGPT ou meta/Llama-2, offrent de nouvelles opportunités pour convertir des questions, formulées en langage naturelle, en langage informatique (génération de code python et/ou SQL). Ces modèles généralistes nécessitent, cependant, d'être ajustés afin de n'extraire que les données réellement attendues par les utilisateurs
	Explicitation sur page suivante

Explicitation du Projet de thèse

1°) Présentation des aspects scientifiques du projet de thèse

Le traitement automatique du langage naturel (TALN) est une discipline qui s'intéresse à la manière dont les ordinateurs peuvent comprendre et traiter le langage humain. Cela inclut la reconnaissance de la parole, la compréhension du sens des mots et des phrases, ainsi que la traduction entre les différentes langues.
Le TALN est un domaine en constante évolution et il existe de nombreuses techniques différentes pour traiter le langage naturel. Certaines de ces techniques sont basées sur des algorithmes de machine learning, tandis que d'autres reposent sur des modèles linguistiques plus formels.
Le TALN est utilisé dans de nombreuses applications, comme les assistants vocaux, les systèmes de traduction automatique, les chatbots et les outils de recherche sur le Web. Il peut également être utilisé dans le traitement de données textuelles, comme l'analyse de sentiments dans les avis de consommateurs ou l'extraction de données à partir de documents juridiques.

Le TALN est un domaine complexe et il y a de nombreux défis à relever pour que les ordinateurs puissent comprendre et traiter le langage humain de manière efficace :

- *La variabilité du langage : Le langage est très variable et il y a de nombreuses façons de dire la même chose. Les ordinateurs doivent être capables de comprendre cette variabilité et de traiter toutes les formes possibles d'une même phrase.*
- *La structure syntaxique et grammaticale : Le langage a une structure complexe qui peut être difficile à analyser pour un ordinateur. Par exemple, les phrases peuvent avoir des structures très différentes et utiliser des verbes et des adjectifs de manière inhabituelle.*
- *Le sens des mots et des phrases : Le sens des mots et des phrases dépend souvent de leur contexte. Les ordinateurs doivent être capables de comprendre ce contexte pour pouvoir déterminer le sens réel des mots et des phrases.*
- *La traduction : La traduction est un défi important pour le TALN, en particulier entre les langues qui ont des structures et des vocabulaires très différents.*
- *Le traitement de données textuelles non structurées : De nombreuses données textuelles sont non structurées, comme les messages instantanés ou les avis de consommateurs. Il est difficile de traiter de telles données de manière automatisée car elles ne suivent pas de structure prédéfinie.*

Ces deux paragraphes ont été générés automatiquement par ChatGPT d'OpenAI, il répondait aux questions :

1. peux-tu me présenter le traitement automatique du langage naturel ?
2. Quels sont les défis du TALN ?

L'intérêt de ce sujet de thèse, dans la continuité d'une collaboration autour de la BDLC, est d'utiliser voire d'adapter les outils actuels au profit du corse qui est une langue peu dotée du point de vue du TAL (Kevers *et al* 2019)[1]. C'est un travail collaboratif car les défis concernent le traitement d'une langue soumise à des phénomènes de variation partiellement décrits et qui, en outre, n'est pas standardisée ce qui constitue un défi pour son traitement. Les derniers développements comme ChatGPT utilisent des « Large Language Model (LLM) » pour saisir le contexte et le sens des mots, c'est l'une des grandes avancées du TALN de 2022 qui faudrait pouvoir appliquer à notre échelle. L'un des objectifs sera de simplifier l'accès aux données de la BDLC en permettant son interrogation en langage naturel, c'est une première étape vers la création d'un chatbot en Corse et la conception d'un outil performent de traduction. Les récentes avancées en intelligence artificielle, dont bénéficient les modèles de langues de type OpenAI/ChatGPT ou meta/Llama-2, offrent de nouvelles opportunités pour convertir des questions, formulées en langage naturelle, en langage informatique (génération de code python et/ou SQL). Ces modèles généralistes nécessitent, cependant, d'être ajustés afin de n'extraire que les données réellement attendues par les utilisateurs

Ce travail de recherche débutera par un état de l'art sur les approches d'apprentissage pour le TALN (machine learning, deep learning, etc.). Cette étape permettra de comprendre les LLM et ainsi voir s'ils sont applicables au vu des ressources actuellement disponibles dans la BDLC. La seconde étape sera d'identifier la meilleure approche d'apprentissage à utiliser en fonction des données à notre disposition et en fonction des avancés de la thèse sur la variation dialectale qui doit débuter en 2023. Enfin, la troisième partie du travail consistera à développer des outils dans la BDLC, afin de valoriser l'apport des outils informatiques et les contenus de la BDLC.

---

[1] L. Kevers, F. Guéniot, A. G. Tognotti, Retali-Medori « Outiller une langue peu dotée grâce au TALN : l'exemple du corse et de la BDLC », TALN-RECITAL, Conférence sur le Traitement automatique des langues naturelles, PFIA 2019, (Toulouse, juillet 2019), <https://www.irit.fr/pfia2019/taln-recital/taln-recital-2019-articles-acceptes/>, pp. 371-380.

2°) Présentation des enjeux de la thèse

Le corse est considéré par l'UNESCO comme une langue « en danger », et qui ne dispose que de peu de ressources et outils dans le monde numérique, et pour le TALN en particulier. A ce titre, elle fait partie des nombreuses langues dites « peu dotées ». Or, il est important d'œuvrer en faveur de la création de l'outillage technologique de la langue car c'est un facteur qui contribue à assurer sa survie et son développement, ce qui correspond à une demande sociale forte dans l'île. Les outils de TALN permettent d'agir sur des leviers clés pour la vitalité d'une langue : le développement des connaissances scientifiques (linguistiques), l'apprentissage, ou encore la réalisation de tâches numériques « quotidiennes ». L'intérêt pour les locuteurs de corse ou toute personne souhaitant investir la langue est donc important.
La recherche participera à l'élaboration d'un ensemble de connaissances et de modules informatiques permettant à terme de constituer une « boîte à outils » de TALN pour le corse.
Cette recherche se place dans la continuité de celles effectuées sur le TALN depuis 2019 au sein de l'UMR CNRS 6240 LISA autour de la Banque de Données Langue Corse (BDLC) qui connaît des développements dans cette perspective depuis lors (Kevers *et al* 2019)[1] (post-doc d'Alice Millour 2022)[2]. Étant donné son orientation informatique importante, elle sera portée par l'UMR CNRS 6134 SPE. La recherche porte sur des aspects importants, tant sur le plan théorique qu'applicatif. Elle pourra donner lieu à une interaction avec les chercheurs déjà actifs sur le sujet, et potentiellement à plusieurs publications de niveau international.

[2] Lemmatisation et annotation morphosyntaxique de corpus dialectaux dans le cadre du traitement automatique des langues (TAL) appliqué au corse » UMR 6240 LISA – Alice Millour 01/01/2022 –31/08/2022