APPEL A PROJETS RECHERCHE 2024

FÉDÉRATION DE RECHERCHE AGORANTIC «CULTURE, PATRIMOINES, SOCIÉTÉS NUMÉRIQUES »

Attention :

- Ne pas dépasser 5 pages
- Transmettre le fichier au format PDF intitulé : ACRONYME-AAP-blc-Agorantic-2024
- Envoyer le fichier à fr-agorantic@univ-avignon.fr **avant le 24 novembre 2023.**

Titre	Modélisation et création d'un corpus pour le résumé de documents textuels en nahuatl (langue mexicaine autochtone) avec des algorithmes d'Intelligence Artificielle
Acronyme	NAHU-AAP-blc-Agorantic-2024
Nom du/des porteur(s)	Juan-Manuel Torres-Moreno (1) Graham Ranger (2) Martha Lorena Avendano Garrido (3) Miguel Figueroa-Saavedra Ruiz (3)
Coordonnées du/de la gestionnaire de laboratoire	Michele Manen - LIA
Laboratoires associés	(1) Laboratoire Informatique d'Avignon (LIA) (2) Laboratoire Identité Culturelle, Textes et Théatrâlité (ICTT) (3) Université de Veracruz, Fac de Mathématiques et Instituto de Investigaciones en Educacion
Budget demandé	8000 Euros
Résumé **Max. 1000 caractères espaces compris**	Nous cherchons à développer des corpus afin d'être employés par des algorithmes de résumés automatiques des documents en nahuatl. Nous voulons également construire un premier système de résumé étalon. Nous nous appuierons sur la combinaison de TAL classique (extraction de phrases pertinentes, analyse superficielle, recherche d'information, corpus ainsi que sur des techniques d'apprentissage profond. Le résumé sera produit directement en nahuatl.

1. Contexte, positionnement, objectif(s)

Actuellement, les politiques visant au maintien et à la revitalisation des langues autochtones minoritaires incluent parmi leurs objectifs l'autonomisation numérique des communautés concernées. Cela répond au fait que ces processus d' autonomisation numérique sont à leur tour liés à leur présence croissante dans l'enseignement supérieur et la recherche et à la diffusion de ces langues comme véhicules de communication et de génération de

connaissances. Ainsi, le développement alphabétisé des communautés et la circulation croissante des textes dans ces langues nécessitent de nouveaux outils, dispositifs et technologies pour leur gestion qui permettent leur traitement pour des processus d'identification, de synthèse et de catalogage dans des bases de données et sur des plateformes d'information pédagogiques, bibliographiques et administratives. Précisément, dans le cadre de la « *Indigenous Languages Decade* » (2022-2032) convoquée par l'UNESCO et plus particulièrement de la « *Declaración de Los Pinos [Chapoltepek]* » — « *Construyendo un Decenio de Acciones para las Lenguas Indígenas* », il est indiqué que « Les technologies numériques jouent un rôle de plus en plus important dans le développement de la société et devraient contribuer à la transmission intergénérationnelle, à la préservation, à la revitalisation et à la promotion des langues autochtones, ainsi qu'à la création dans ces langues ». Bien que cet objectif ne soit pas encore réalisable dans une bonne partie des langues nationales parlées au Mexique, la langue nahuatl, possède une culture écrite continue. Par ailleurs, de manière plus récente, elle est de plus en plus utilisée pour la production de textes académiques (thèses, manuels, articles et livres scientifiques) qui, compte tenu du développement de l'alphabétisation académique et de la production de textes sous forme numérique, génèrent des besoins d'archivage (enregistrement, classification, organisation) pour une meilleure diffusion auprès des publics. Pour cette raison, ce défi, qui a en soi un impact sociolinguistique, implique également une opportunité de progrès dans le développement d'applications statistiques et informatiques dans le cadre de projets de linguistique appliquée et computationnelle qui peuvent trouver une étude de cas pertinente et significatif dans le nahuatl.

Dans un autre coté, on sait que les algorithmes d'apprentissage profond (IA) sont de plus en plus utilisés pour créer des textes artificiels. Dan ce projet nous nous proposons d'utiliser des techniques d'IA pour augmenter la démocratisation des technologies numériques. D'où l'intérêt majeur de la dimension interdisciplinaire, à la fois théorique et appliquée de ce projet. Le premier objectif du projet consiste d'abord à modéliser les textes en nahuatl pour les représenter dans un espace mathématique approprié. Le second objectif radicalement innovant du projet est le suivant : formaliser le fonctionnement et les impacts d'un système de résumé automatique (paragraphes, phrases, morceaux de texte) en prenant appui sur plusieurs domaines d'activité en Informatique : l'apprentissage profond, le traitement automatique des langues (TAL), l'optimisation et la théorie des graphes. Ce système devrait permettre, grâce à ces algorithmes appropriés, de proposer des résumés originaux et adaptés aux besoins spécifiques des utilisateurs. Il devra également tenir compte de un ensemble des contraintes proposées par l'utilisateur. Enfin, il est prévu d'évaluer le système, ainsi que d'expliquer le pourquoi de telle ou telle réalisation textuelle. Ce projet pourra devenir l'étendue des travaux précédents sur le même domaine, qui se limitaient à la génération de résumés en français, anglais et espagnol, car le travail sur le nahuatl viendra élargir la portée des nos algorithmes.

2. Questionnement scientifique.

Cette projet abordera plusieurs problèmes, dont la résolution potentielle permettra de lever des verrous scientifiques pluridisciplinaires importants :

i/ Génération des corpus ad hoc : il faut une compilation (automatique et/ou manuelle) de corpus adéquats permettant d'anticiper des structures textuelles dans la langue – nahuatl français, espagnol – choisie. La problématique scientifique concerne entre autres, l'étude et

analyse des corpus et des outils d'analyse linguistique pour le nahuatl.

ii/ Représentation textuelle: il faudra créer des représentations abstraites adéquates (sac de mots, plongements de mots, graphes, etc.) pour saisir l'informativité contenue dans les structures grammaticales textuelles. La problématique scientifique abordée concerne les méthodes d'apprentissage profond, linguistiques et l'allocation de ressources dans un réseau textuel qui doit montrer cohésion et cohérence dans son argumentation et sa sémantique.

iii/ Génération des résumés: La problématique scientifique concerne la fouille de textes et plus précisément, la production automatique de résumé textuelle par extraction. D'abord les résumés seront produits dans un contexte monolingue (nahuatl) et dans une deuxième phase nous allons produire probablement des résumés bilingues (nahuatl-français, nahuatl-espagnol) ou trilingues.

Dans ce projet exploratoire, nous allons nous concentrer sur le verrou scientifique i/, qui concerne l'analyse linguistique superficielle du nahuatl et la création de corpus bilingues (fr-nahuatl, es-nahuatl).

3. Méthodologie

Nous utiliserons de outils de Traitement Automatique de Langues (TAL) que le LIA maîtrise depuis longtemps, afin de pouvoir constituer et traiter des corpus issus des documents, littéraires ou pas, en nahuatl et en espagnol. Ces algorithmes ont déjà obtenu de bons résultats dans des tâches telles que la classification automatique, le résumé automatique, la détection et la classification d'opinions. Ces outils restent assez indépendants de la langue et de la thématique, et sont par conséquent facilement adaptables aux besoins des expériences proposées. Après avoir mis au point des méthodologies opérationnelles permettant de modéliser les textes en nahuatl, il s'agira de passer à développer un résumeur étalon (*baseline*) en nahuatl et à l'analyse des résultats. Enfin, il s'agira, en faisant travailler ensemble des méthodologies diverses, de se situer dans un champ de recherche où très peu de travaux se sont penchés sur ces thématiques, et encore moins dans une telle perspective d'étude du nahuatl, où nous voulons démontrer que ces méthodes sont appropriées pour traiter cette langue très éloignée des langues indo-européennes.

4. Résultats attendus et caractère innovant de la recherche

D'abord nous voulons constituer des corpus bien caractérisés qui seront probablement alignés (nahuatl-espagnol, nahuatl-français) et constitués de texte source-texte résumé. Ils seront de taille petite. Ces corpus ne seront pas utilisées pour apprentissage mais dans le cadre d'une extraction non supervisée de phrases porteuses d'information. L'évaluation des résultats se fera du point du vue quantitatif et qualitatif. En ce qui concerne l'évaluation quantitative, des statistiques et une analyse fine au niveau des algorithmes TAL et des résultats en termes de Recherche d'information, seront établis : précision, rappel, F-score. Également, des statistiques au niveau des n-grammes (ROUGE et calculs des divergences de distribution des probabilités) seront effectués. Le terrain de recherche, la méthodologie et les corpus en nahuatl serviront de base à l'évaluation qualitative. L'évaluation qualitative se fera au moyen d'une lecture directe faite par un expert dans la langue nahuatl. Pour cela, nos collègues de l'Universidad Veracruzana (Instituto de Investigacion en Educacion) vont nous apporter une aide précieuse car ils disposent des ressources humaines et linguistiques nécessaires. Une attention importante sera portée aux publications dans des congrès scientifiques nationales et internationales.

5. **Dimension interdisciplinaire**

Notre projet s'inscrit fortement dans les axes suivants : 2-Culture et numérique et 5-Structuration et exploitation de corpus (SEC). Il est axé sur le TAL, la linguistique computationnelle et les corpus.

6. **Partenariats extérieurs envisagés**

Nous allons établir des collaborations et/ou conventions avec l'Universidad Veracruzana (Facultad de Matematicas et l'Instituto de Investigaciones en Educacion) afin de réaliser nos recherches dans les meilleures conditions.

7. **Objectifs de pérennisation du projet**

Nous allons probablement demander une subvention interna de l'Universidad Veracruzana pour compléter les mobilités internationales.

8. **Expression des besoins en assistance informatique**

Rien n'est prévu pour l'instant.

9. Budget (€) prévisionnel *

	Brève description	**Montant**
Missions	2 missions Mexique<->France	4000
Consommables, petits matériels**		
Organisation de réunions		
Stages*, **vacations**	3 mois de stage master	1800
Prestations de service	Création des corpus	2200
Budget total		8600
Co financements le cas échéant		
Budget demandé à Agorantic		**8000**
Recettes extérieures	**Demande à UV en cours**	**600**

* Veuillez modifier les catégories de dépenses si besoin – ajoutez/supprimez des lignes à votre convenance

** Petit matériel ne dépassant pas les 600€

**Gratification de stage obligatoire au-delà de 2 mois – prévoir environ 600€ par mois

NB : si le projet a déjà fait l'objet d'un financement lors d'un précédent AAP : justifier la nouvelle demande et présenter les évolutions du projet.