# PROPOSITION SUJETS DE THESE
# CONTRATS DOCTORAUX
# 2024-2027

**X Contrat doctoral fléché FR Agorantic**
**X Contrat doctoral fléché EUR InterMEDIUS**
→Dans le cas où vous souhaitez candidater sur les deux contrats doctoraux, veuillez cochez les deux cases.

| | | | |
|---|---|---|---|
| **Directeur de thèse** : | **Juan-Manuel Torres-Moreno, HDR HC** | Mail: | **juan-manuel.torres@univ-avignon.fr** |
| Laboratoire : | **Laboratoire Informatique d'Avignon (LIA)** | Téléphone: | **0490843568** |
| **Co-directeur.rice et/ou encadrant.e** : | **Graham Ranger (Co-directeur)[1]<br>Martha Lorena Avendaño Garrido (encadrante)[2]<br>Miguel Figueroa-Saavedra Ruiz (encadrant)[3]** | | |
| Laboratoire: | **[1]Laboratoire Identité Culturelle, Textes et Théatrâlité (ICTT)<br>[2]Facultad de Matemáticas (UV Mexique)<br>[3]Instituto de Investigación en Educación (UV Mexique)** | | |

**Titre en français** : Abstraction et synthèse de documents textuels en nahuatl (langue mexicaine autochtone) en utilisant des algorithmes d'Intelligence Artificielle

**Titre en anglais** : Abstraction and synthesis of textual documents in Nahuatl (indigenous Mexican language) using Artificial Intelligence algorithms

**Résumé en 7 lignes** : Nous cherchons à développer des algorithmes de résumés automatiques des documents en nahuatl. Nous nous appuierons sur la combinaison de TAL classique tel que l'extraction de phrases pertinentes, analyse superficielle, recherche d'information, corpus ainsi que sur la production de texte au moyen de techniques d'apprentissage profond (RN du type transformateur). Le résumé sera produit directement en nahuatl et probablement traduit en français et en espagnol.

**Mots clés : Traitement Automatique des Langues ; Nahuatl ; Corpus ; Résumé automatique de textes ; IA ; Optimisation**

# 1- Présentation du sujet (3 pages maximum):

**Objectif scientifique**

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 (voir https://www.unesco.org/fr/taxonomy/term/423821) et plus particulièrement de la « *Declaración de Los Pinos [Chapoltepek]* » — « *Construyendo un Decenio de Acciones para las Lenguas Indígenas* » : (voir le lien https://unesdoc.unesco.org/ark:/48223/pf0000374030_spa), 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 » (p. 9). 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 cette thèse 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 cette thèse. 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 envisagé d'évaluer le système, ainsi que d'expliquer le pourquoi de telle ou telle réalisation textuelle. Cette thèse pourra devenir l'étendue des travaux précédents sur le même domaine, qui se limitaient à la génération de résumés en français, anglais et espagnol (7,8,12,13), car le travail sur le nahuatl viendra élargir la portée des nos algorithmes.

**Verrous scientifiques**

Cette thèse 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 ou manuelle) de corpus adéquats (6,10,11,13) permettant d'anticiper des structures textuelles dans la langue – nahuatl français, espagnol – choisie. Un projet exploratoire concernent cette problématique vient d'être soumis à Agorantic/Intermedius. 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 (1)

**Cadre expérimental et originalité du sujet**

Nous utiliserons de outils de Traitement Automatique de Langues (TAL) que le LIA maîtrise depuis longtemps (2, 3, 4) , 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 (4). Ces outils restent assez indépendants de la langue et de la thématique, et sont par conséquent facilement adaptables aux besoins des expériences proposées. Nous utiliserons également des outils d'optimisation et des algorithmes gloutons pour générer des solutions approximées à ces problèmes. Ces méthodes sont déjà utilisées dans les réseaux de contenus pour la diffusion de vidéos (2). Après avoir mis au point des méthodologies opérationnelles permettant de modéliser les textes en nahuatl, il s'agira de passer à l'analyse des résultats, du point de vue sémantique. Une originalité de ce projet, consistera à étudier le phénomène (peu connu) d'hallucination des réseaux profonds dédiées au traitement textuel (le fait d'introduire des passages de texte incohérent ou éloigné de la sémantique prévue) (3). Il faut étudier, comprendre et contrôler ce phénomène, pou évite le biais dans nos algorithmes de résumé. 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.

# Évaluation

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 https://www.uv.mx/iie) vont nous apporter une aide précieuse car ils disposent des ressources humaines et linguistiques

nécessaires. Une attention sera portée aux publications du doctorant-e qui, pendant la durée du contrat doctoral, sera encouragé à publier dans des revues scientifiques nationales et internationales.

## 2- Profil du/de la candidat(e): *Nous utilisons « candidat » pour alléger le texte.*

Pour cette recherche, nous recherchons des candidats ayant une certaine expérience en programmation et algorithmique. Il est intéressant que le candidat sache programmer dans au moins 2 de langages suivants : C/C++, python, perl, ruby, R. Également il doit être à l'aise dans la programmation en bash sous environnement GNU/Linux. Nous souhaitons également que le candidat ait ou soit disposé à étudier des algorithmes, packages et techniques d'intelligence artificielle et apprentissage profond, ainsi qu'il ait des connaissances ou ait suivi des cours en mathématiques (algèbre linéaire et optimisation principalement). Puisque la production de texte sera trilingue (français, espagnol, nahuatl), la connaissance de la langue espagnole sera un atout.

## 3- Opportunités de mobilité à l'international du/de la doctorant(e) dans le cadre de sa thèse:

Oui, très probablement à la Faculté de Mathématiques de l'Universidad Veracruzana (Mexique) si le financement du séjour est acquis. Au moins deux court séjours sont prévus.

## 4- Références bibliographiques

(1) J-M Torres-Moreno, Automatic Text Summarization, Wiley, London 2014

(2) Juan-Manuel Torres-Moreno. Résumé automatique de documents : Une approche statistique. Hermès Lavoisier, 2011, ISBN 978-2-7462-3212-9

(3) Abstraction ou hallucination ? Etat des lieux et évaluation du risque pour les modèles de génération de résumés automatiques de type séquence-à-séquence, Akani, Eunice, Favre, Benoit and Bechet, Frederic, Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1, 6, 2022, Avignon, France

(4) Juan-Manuel Torres-Moreno, Marc El-Bèze, Frédéric Béchet, Nathalie Camelin, Fusion probabiliste appliquée à la détection et classification d'opinions, DEFT'09, Paris, France, 22 juin 2009, 15p.

(5) G Bel-Enguix, G Sierra, H Gómez-Adorno, J-M Torres-Moreno, J-G Ortiz-Barajas, J Vásquez. Overview of PAR-MEX at Iberlef 2022: Paraphrase Detection in Spanish Shared Task. Procesamiento de Lenguaje Natural (PLN),

(6) L Moreno, J-M Torres-Moreno. LiSSS: a new multi-annotated multi-emotional corpus of Literary Spanish Sentences, CyS, 24(3) :1139–1147, 20203

(7) L Moreno, J-M Torres-Moreno, R Wedemann. Generación de frases literarias: un experimento preliminar, PLN, 65 :29–36, 2020

(8) L Moreno, J-M Torres-Moreno, E SanJuan, R Wedemman. Automatic Generation of Literary Sentences, Linguamática, 12(1) :15-30, 2020 5

(9) L Moreno, J-M Torres-Moreno, C González. Estudio de hiperparámetros de modelos neuronales en la generación de frases literarias. Research in Computing Science (RCS), 150(5), 2021 20

(10) I Morgado, L Moreno, J-M Torres-Moreno, R Wedemann. MegaLite-PT: A Corpus of Literature in Portuguese for NLP. BRACIS 2022. Intelligent Systems pp 251–265

(11) L Moreno, J-M Torres-Moreno. MegaLite-2 : An Extended Bilingual Comparative Literary Corpus. Computing Conference 2021. pp 1014-1029, 2022

(12) L Moreno, J-M Torres-Moreno, R Wedemann. A Preliminary Study for Literary Rhyme Generation based on Neuronal Representation, Semantics Resources and Shallow Parsing. STIL 2021. pp 190–198

(13) L Moreno, J-M Torres-Moreno. Megalite : A New Spanish Literature Corpus for NLP Tasks. 8th International Conference on Artificial Intelligence and Applications (AIAP'21), pp. 131-147, 2021

(14) Avendaño-Garrido M.L., Gabriel-Argüelles J.R., Torres-Quintana L. and González-Hernández J. (2018) An approximation scheme for the Kantorovich-Rubinstein problem on compact spaces Journal of Numerical Mathematics.

(15) Avendaño-Garrido M.L., Gabriel-Argüelles J.R., Mezura-Montes E. and Quintana-Torres L. (2016). A metaheuristic for a numerical approximation to the Mass Transfer problem. International Journal of Applied Mathematics and Computer Science.

(16) Figueroa-Saavedra, M. (2023). Marcadores y conectores discursivos en la textualidad náhuatl entre universitarios nahuahablantes. Cultura, Lenguaje y Representación, 31, 237–263. https://doi.org/10.6035/clr.6816

(17) Figueroa-Saavedra, M., Bernal-Lorenzo, D. et Nava Vite, R. (2022). In tlahkuilolyotl ken se nawatlahtolchikawalistli ipan weyitlamachtiloyan: se tlachiwalistli tlen moneki axkan mochiwa. CPU-e, Revista de Investigación Educativa 35, 91-120.

(18) Mager, M., Gutiérrez-Vásques, X., Sierra, G.et Meza, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), pp. 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

(19) Figueroa-Saavedra, M., Alarcón-Fuentes, D., Bernal-Loenzo, D. et Hernández-Martínez, J.A. (2014). The incorporation of national indigenous languages into the academic development of universities: The experience of the Universidad Veracruzana. Revista de la Educación Superior 43(171), 67-92

**x J'ai informé le Directeur de mon unité du dépôt de cette proposition de sujet de thèse**

Les sujets devront être adressés avant le 11 décembre 2023 midi aux adresses agorantic@univ-avignon.fr et intermedius@univ-avignon.fr

Maximum 5 pages (titre fichier : NAHU-LIA-TORRES-AGORANTIC.InterMEDIUS)