PROPOSITION SUJETS DE THESE
CONTRATS DOCTORAUX FR AGORANTIC 2026—2029

Directeur de thèse :	Juan-Manuel Torres	**Mail :**	juan-manuel.torres@univ-avignon.fr
Laboratoire porteur :	LIA	**Téléphone :**	0490843568
Co-directeur et co-enca-drant et leurs e-mails :	Graham.Ranger: graham.ranger@univ-avignon.fr; Miguel Figueroa-Saavedra : migfigueroa@uv.mx		
Laboratoires associés :	LIA/ICTT (Avignon Université), Universidad Veracruzana (Mexique)		

Traduction interdialectale en nawatl, langue autochtone du Mexique

Interdialectal translation into Nahuatl, an indigenous language of Mexico

Nous développerons des algorithmes pour la traduction interdialectale entre variétés du nawatl (langue autochtone nationale du Mexique). En effet, les textes disponibles dans cette π-langue (peu dotée de ressources informatisées), sont rares et la diversité linguistique est très grande. Également, leurs diverses graphies compliquent plus la tâche. Dans au moins 4 régions (Centrale, Nord, Pacifique, Golfe) le nawatl a trop évolué (les vocabulaires différent) et leurs caractéristiques grammaticales sont distinctes. La langue est complexe au niveau grammatical, agglutinante et polysynthétique et les ressources informatisées sont pratiquement inexistantes. Un système interdialectal avec des modèles IA profonds aura un impact social car il permettra de mieux communiquer entre les communautés nahuas.

Mots clés : Nawatl ; Langues pi ; TAL ; Traduction automatisée ; Traduction interdalectale ; IA ; Transformateurs

---

1. **Disciplines concernées :** Informatique ; TAL ; Apprentissage profond ; Linguistique computationnelle
2. **Présentation du sujet**

Contexte, positionnement, objectifs : La langue nawatl [3,4], une des langues nationales du Mexique, parlée par environ 2,5 millions de personnes (Amérique du Nord et Centrale) compte 30 variétés dialectales selon l'INALI (Institut national des langues indigènes). Le nawatl est de plus en plus utilisé 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

et organisation pour une meilleure diffusion. Pour cette raison, ce défi socio-linguistique, implique une opportunité pour développer des outils informatiques intéressants pour le public (nawaphone ou pas). Or, cette diversité graphique et dialectale pose des problèmes importants pour leur traitement automatisé et même pour les personnes [6]. Cette diversité, qui témoigne de l'évolution historique et culturelle des communautés parlant la langue nawatl au cours des deux derniers siècles, a été considérée comme un facteur que les locuteurs eux-mêmes ont perçu comme une faiblesse. Cela est d'autant plus flagrant qu'il n'existe pas de variété standard reconnue du nawatl, ce qui conduit à considérer l'espagnol comme langue pivot. Cela ne favorise pas la possibilité de nouveaux processus de standardisation et d'apprentissage des variations linguistiques en tant que communautés linguistiques. Dans le cas de la communication écrite, ce phénomène est encore plus marqué. En plus, la rareté des corpus des langues autochtones ou π-langues [1,2] vient encore s'ajouter à ces difficultés. En raison des problèmes évoqués, il n'est pas simple de constituer des corpus ayant des bonnes propriétés (en taille et qualité) pour l'apprentissage automatique. En effet, notre projet NAHU[2] a servit à la compréhension basique du nawatl [4]; au développement des algorithmes de compression des embeddings et à la constitution d'un nouveau corpus nawatl, π-yalli [5]. Le nawatl étant une langue agglutinante et polysynthétique, nous utilisons des représentations de mots et tokens afin de pouvoir extraire des caractéristiques sémantiques fondamentales (la racine des verbes et leurs conjugaisons). Dans ce projet nous nous proposons d'utiliser des techniques d'IA et de linguistique computationnelle pour créer un traducteur automatique interdialectal (TAI). La traduction automatique (TA) est un problème qui consiste à automatiser la tâche de traduction d'une phrase vers une autre langue cible. Les caractéristiques distinctives des langues autochtones [10], telles que la morphologie polysynthé-tique, les variations morphologiques importantes et l'orthographe non standardi-sée, posent des défis particuliers aux modèles de TA qui reposent sur la correspon-dance exacte au niveau lexical ou de caractères, en particulier lorsque ces mesures n'ont pas été spécifiquement testées dans ces langues [13]. D'autre part, les modèles IA basés sur des LLM (Mistral, Gemini, etc.), nécessitent un volume trop important de données [9] pour saisir la représentation linguistique sous-jacente, ce qui n'est pas disponible pour les π-langues. Près de la moitié des 7000 langues parlées dans le monde sont actuellement menacées. Les experts prévoient que près de 1500 d'entre elles pourraient disparaître d'ici la fin du siècle en raison de plusieurs facteurs (mondialisation, croissance économique, soutien insuffisant accordé aux π-langues) [7]. Les langues autochtones ne sont pas seulement des joyaux culturels, elles recèlent également des perspectives et une cosmovision unique. La TA de ces langues représente un défi de taille en raison de la rareté des ressources numériques et des corpus parallèles. Cependant, le nawatl a fait l'objet de quelques études comme la TA statistique (SMT) et la TA neuronale (NMT), dans des réseaux neuronaux récurrents (RNN) [11,12]. Mais il n'y a aucune étude de TAI entre les principales variétés

dialectales nawatl, à notre connaissance. Au niveau phraséologique, voici un exemple de la diversité et de la complexité évoquées, avec la phrase : « ***Il y avait un homme marié qui avait une femme*** » qui peut se traduire par :

Occidental	*Niman kataya se lakal munamiktijtuk kipiataya isiua*
Central	*Melak yokatka sentetl tlakatl monamiktitok kipiaya un isowah*
	Nelli katki se tlakatl monamiktihtok okipiyaya in isiwaw
Huasteca	*Nelia itstoya se tlakatl kipiayaya ni isiwaj*
Oriental	*Onoya se tagat kipiaya monamiktitok ipalmiya*
	Nemik se takat munamiktijtuk kipiatuya ne isiwaw

Les objectifs du projet consistent d'abord, à créer un classifieur de dialectes. Ensuite, à développer un segmenteur (tokeniseur) efficace de mots nawatl, et puis à développer des prototypes d'un ***Traducteur Automatique Interdialectal*** (TAI).

Questionnement scientifique : Ce projet de thèse abordera plusieurs problè-mes, dont leur résolution permettra de lever des verrous scientifiques pluridisci-plinaires importants : **i/ Identification automatique des variétés nawatl** : La problé-matique scientifique concerne la détection via des classifieurs statistiques et neuronaux [14,15,16] des variétés dialectales nawatl. **ii/ Segmenteur de mots nawatl** (*tokeniseur*) **:** La problématique concerne la segmentation automatique (par apprentissage et avec des règles linguistiques) des mots (agglutinés et poly-synthétiques) de textes venant des variétés dialectales. **iii/ Traducteur inter-dialectal nawatl :** à l'aide d'apprentissage automatique et profond sur des corpus nawatl (allignés ou pas) et des reglès linguistiques nous développerons un TAI interdialectal.

Tout cela soulève certaines questions scientifiques : Quels transformateurs sont les meilleurs, les plus adaptés ou à utiliser dans le cas du nawatl ? Quel genre de tokeniseurs utiliser ? La classification dialectale peut aider à développer d'autres genre de traducteurs ? Il vaut mieux s'appuyer sur une langue pivot pour bien saisir la sémantique sous-jacente ? Ce sont toutes des questions ouvertes auxquelles nous essaierons de répondre dans ce projet de thèse.

Méthodologie : Nous utiliserons des outils TAL et d'apprentissage profond [9] afin de pouvoir constituer, tout d'abord un classifieur dialectal (n-grammes et neuronal), puis un segmenteur de mots, qu'il soit statistique, basé sur des règles, par appren-tissage ou leur combinaison. Par la suite, nous allons concevoir et développer un traducteur interdialectal basé sur des représentations riches de mots et transforma-teurs. Ces trois tâches sont toutes originales à notre connaissance. Nous allons combiner la puissance des méthodes de apprentissage profond et la simplicité des méthodes TAL pour cela. En fin, des évaluations (quantitatives et qualitatives) s'avè-rent indispensables pour mesurer les résultats produits par nos systèmes.

Résultats attendus, caractère innovant de la recherche : Nous comptons

identifier correctement les différentes variétés du nawatl. Puis, grâce au segmenteur de mots, nous espérons identifier les morphèmes locatifs, temporels, de négation, révérence, les diminutifs et les racines verbales qui, dans de nombreux cas, ne changent pas entre variétés. En ce qui concerne le TAI, nous espérons que les modèles d'apprentissage profond, compte tenu de leur capacité d'abstraction et de synthèse, permettront de générer une variété nawatl homogène et exploitable. Tandis que pour l'évaluation quantitative, des statistiques et une analyse fine se feront en utilisant des mesures état de l'art (adaptées au nawatl) telles que BLUE, METEO et TER [11-12]. L'évaluation qualitative se fera au moyen de lecture et d'analyses directes faites par des experts nawaphones (nos collègues de l'Université Veracruzana apportent déjà un soutient logistique important car ils disposent des ressources humaines et linguistiques nécessaires). Les résultats du projet, typiquement les corpus aligné dialectal et nawatl français, seront disséminés largement par le biais d'une application en ligne de concordancier cqpweb [8], dont une instance est hébergée sur les serveurs de l'université, et qui fonctionnent déjà pour la mise en ligne des corpus authentiques, avec et sans annotations POS et lexicales. Nos résultats obtenus permettront d'enrichir la mise en ligne des données nawatl (https://demo-lia.univ-avignon.fr/pi-yalli), avec la possibilité d'y intégrer des métadonnées et les modèles résultants pour la diffusion auprès des communauté scientifique et nawaphone. C'est surtout dans cette dernière qu'on aura un impact sociétale pour renforcer la compréhension mutuelle et le rapprochement des variétés vers des formes qui permettent de mieux communiquer. Une attention particulière sera portée aux publications dans congrès scientifiques et dans de revues (nationales et internationales) des résultats obtenus sur l'ensemble de la thèse.

Dimension interdisciplinaire : Notre projet exploratoire s'inscrit fortement dans la description, modélisation, théorisation linguistique, TAL sur les π-langues, corpus (leur constitution, exploitation et valorisation) et l'apprentissage profond (IA).

Pérennisation : Ce projet de thèse présente plusieurs difficultés majeurs; d'où la nécessité de le pérenniser, pour bien boucler la traduction interdialectale automatisée et la traduction entre le nawatl et langues bien dotées en ressources informatisées, au moyen d'un AAP bi/internationale Mexique/(France/Canada).

3. Profil du/de la candidat·e.

Pour ce projet nous recherchons des candidat-e-s ayant une certaine expérience en programmation et algorithmique. Il est intéressant que le candidat-e sache programmer dans au moins 2 de langages suivants : C/C++, python, perl, ruby, prolog, docker. Également elle/il doit être à l'aise dans la programmation bash sur GNU/Linux. Nous souhaitons également que le candidat-e soit disposé-e à étudier des algorithmes, packages et techniques d'intelligence artificielle et apprentissage profond. Il faut des connaissances en mathématiques (algèbre linéaire et optimisation principalement). La connaissance des langues nawatl et espagnol seront un vrai atout.

4. Opportunités de mobilité à l'international du/de la doctorant·e dans le cadre de sa thèse.

Nous avons déjà établit des collaborations concrètes au Mexique avec l'Université Veracruzana. Mme Avendano et M Figueroa co-dirigent la thèse de M Guzmán portant sur l'analyse automatisée de nawatl. Par ailleurs, nous réalisons des recherches conjointes avec l'UNAM et Colmex sur le domaine, et nous allons poursuivre et approfondir ces études. Le candidat-e sera amené-e à faire au moins deux séjours de recherche -d'un mois- au Mexique.

5. Références bibliographiques

(1) Abdillahi, N., et al. (2006). Boites à outils TAL pour les langues peu informatisées : le cas du Somali. JADT, Besançon.

(2) Berment, V. (2004). Méthodes pour informatiser les langues et les groupes de langues "peu dotées". Thesis Univ Joseph-Fourier - Grenoble I.

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

(4) Figueroa, M., et al. (2014). The incorporation of national indigenous languages into the academic development of universities: The experience of the UV. Revista de Educación Superior 43(171), 67-92.

(5) Guzmán, J.J., et al (2025). π-yalli: un nouveau corpus pour le nahuatl. TALN Marseille, pp 802–816. ATALA.

(6) Guzmán, J.J., et al. (2025). A symbolic algorithm for the unification of nawatl word spellings. Advances in Soft Computing, MICAI'25, pp 141–154. SMIA.

(7) Hansen, M.P. (2024). Nahuatl Nations: Language Revitalization and Semiotic Sovereignty in Indigenous Mexico. Oxford Univ Press.

(8) Hardie, A., (2012) CQPweb-Combining Power, Flexibility and Usability in a Corpus Analysis Tool. IJCL, v17(3):380-409.

(9) L Moreno, J-M Torres-Moreno. Megalite : A New Spanish Literature Corpus for NLP Tasks. 8th AIAP'21, pp. 131-147, 2021

(10) Mager, M. et al. (2018). Challenges of language technologies for the indigenous languages of the Americas. 27th COLING'18, pp. 55–69, Santa Fe, New Mexico, USA. ACL.

(11) Bello García, et al. (2021). Nahuatl Neural Machine Translation Using Attention Based Architectures: A Comparative Analysis for RNNs and Transformers as a Mobile Application Service. MICAI.

(12) M Yahan and M Islam. 2025. Leveraging Large Language Models for Spanish-Indigenous Language Machine Translation at AmericasNLP 2025. 5th Workshop, pp 126–133, Albuquerque, ACL.

(13) N Krasner, et al. 2025. Machine Translation Metrics for Indigenous Languages Using Fine-tuned Semantic Embeddings. In Indigenous Languages of the Americas (AmericasNLP), pp 100-104, Albuquerque, ACL

(14) Cavnar, W B, & Trenkle, J M (1994). N-gram-based text categorization.

(15) C Zhou et al. 2015. A c-lstm neural network for text classification. arXiv:1511.08630.

(16) Y Kim. 2014. Convolutional Neural Networks for Sentence Classification. EMNLP, pp1746–1751, Doha, Qatar. ACL.

X J'ai informé le directeur de mon unité du dépôt de cette proposition de sujet de thèse