APPEL A PROJETS RECHERCHE 2025

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-2025
- Envoyer le fichier à fr-agorantic@univ-avignon.fr **avant le 10 mai 2024.**

Titre	Unification des graphies des documents textuels en nahuatl et leur modélisation en utilisant l'IA
Acronyme	**NAHU²**
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	Michèle MANEN Laboratoire Informatique d'Avignon (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 €
Résumé **Max. 1 000 caractères espaces compris**	Nous cherchons à développer des algorithmes pour unifier les graphies des documents en nahuatl venant à la fois : a/ de sources hétérogènes ; b/ de diverses propositions de graphies et c/ de variantes linguistiques régionales. En effet, les documents sont disponibles en plusieurs formats et codification numériques distinctes (PDF, texte, OCR, utf, isolatin). Également, il y a plusieurs alphabets proposés : franciscain, jésuite, traditionnel, OPINAC, SLI et pratique, et actuellement l'INALI tente de créer une norme orthographique unifiée et aussi l'Univ. Veracruzana dans ses propes processus éducatifs et administratifs. Enfin, il y a au moins 4 régions (Centrale, Nord, Pacifique, Golfe du Mexique) où le nahuatl a évolué et suivi des chemins différents (les vocabulaires ne sont pas les mêmes, et leurs caractéristiques grammaticales distinctes bien qu'ils partagent une structure grammaticale commune, ce qui complique le développement de corpus adéquats pour un traitement informatique. Nous voulons construire un système d'unification des graphies, pré-traitement indispensable pour la constitution de corpus. Nous nous appuierons sur la combinaison de TAL, RI, ainsi que sur l'IA profonde.

1. Contexte, positionnement, objectif(s)

De nos jours, 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. De manière plus récente, la langue nahuatl 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. Toute cette culture écrite spécialisée nécessite des outils d'IA qui permettent la localisation et la gestion des textes, en surmontant la diversité graphique et dialectale, ce qui permettra la génération de corpus textuels, la synthèse d'informations et l'apprentissage automatique.

En effet, en nahuatl on utilise de nombreux digraphes, des lettres composées (TZ, TL, HU/UH, CU/UC), qui sont realisés avec des lettres utilisées aussi de façon isolée, et qui peuvent être confondues lorsqu'elles sont juxtaposées, ce qui rend difficile la différenciation des syllabes. Cela arrive surtout avec CH qui n'est pas égal à C-H (exemple : tla-chia pas tlac-hia / tlac-hua, pas tlach-ua. Ici la question ne vient pas des lettres, mais des syllabes et cela changera en fonction de la norme orthographique, car certaines de ces lettres dans d'autres alphabets ont été remplacées par une seule lettre, par exemple : HU-→W. Or, cette diversité graphique et dialectale pose des problèmes importants pour leur traitement automatisés (et même pour les personnes). La rareté des corpus vient encore s'ajouter à ces difficultés. En raison des problèmes évoqués, il n'est pas simple de générer des corpus avec des bonnes propriétés (en taille et qualité) pour l'apprentissage automatique. En effet, le projet NAHU a servit, entre autres, à la compréhension basique du nahuatl (suivit des cours en ligne, exercices, etc) (Figueroa-Saavedra et al 2014, 2020-24); au développement des algorithmes de compression dimensionnelle des embeddings (Avendano-Garrido et al, 2024). Le nahuatl étant une langue agglutinante, nous aurons besoin des embeddings de mots et de caractères (Watson, 2018), afin de pouvoir extraire les caractéristiques sémantiques fondamentales (en particulier la racine des verbes et leurs conjugaisons). Or, les embeddings de caractères sont bien plus volumineux que ceux de mots. D'où le besoin de les compresser de façon adéquate. Le travail précédent (Avendano-Garrido et al, 2024) nous permettra de diminuer cette dimensionnalité qui peut devenir gênante. Dans ce projet nous nous proposons d'utiliser des techniques d'IA pour constituer des corpus en nahuatl adéquats à l'apprentissage profond. Nous ignorons pour l'instant quelle est la taille minimale pour un apprentissage correcte des texte en nahuatl avec des méthodes d'IA. Combien de textes, de mots ou de caractères sont-ils nécessaires ? Quelles méthodes (statiques ou transformers) sont les meilleurs ou les plus simples à adapter ou à utiliser dans le cas du nahuatl ? Ce sont des questions ouvertes auxquelles nous essaierons de répondre. 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. NAHU² sera en mesure de continuer avec le projet précédent dans la partie résumé automatique (déjà en cours de réalisation avec un système de résumé par graphes -CAFETAL- et embeddings, pour le moment en français). Les algorithmes d'apprentissage automatique développés dans NAHU² pourront produire les embeddings en nahuatl dont CAFETAL a besoin pour la production directe de résumés en nahuatl. Enfin, il est prévu d'évaluer les systèmes générés dans ce projet.

2. Questionnement scientifique

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

i/ Standardisation des documents textuels en nahuatl pour unifier les graphies (en particulier suivant celle de *Instituto Nacional de Lenguas Indígenas, INALI*), ayant comme objectif la génération de corpus adéquats.

Ii/ Génération des corpus : 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 le projet précédent pour l'étude et analyse des corpus et des outils d'analyse linguistiques.

iii/ Génération des résumés: La problématique scientifique concerne la fouille de textes et plus précisément, la production automatique de résumé textuelle par extraction. Les résumés seront produits dans un contexte monolingue (nahuatl). Ce verrou fait partie du projet précédent.

Dans ce projet nous allons nous concentrer sur les verrous scientifiques i/ et ii/, qui concerne l'analyse linguistique superficielle du nahuatl et la création de corpus bilingues (français-nahuatl, espagnol-nahuatl).

3. Méthodologie

Nous utiliserons de outils de Traitement Automatique de Langues (TAL) que le LIA maîtrise depuis longtemps (Torres et al, 2009), afin de pouvoir constituer et traiter des corpus issus des documents en nahuatl. Ces outils restent assez indépendants de la langue et de la thématique, et sont par conséquent facilement adaptables aux besoins des expériences proposées. Nous nous proposons d'étudier plusieurs représentations d'embeddings de mots et des caractères ainsi que leur réduction dimensionnelle afin d'en trouver la meilleure pour les objectifs du projet.

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.

Les méthodes utilisées concernant à la fois des algorithmes de :

- Recherche d'Information (RI) (Gaussier, 2013, Torres-Moreno at al 2015) ;
- Traitement Automatique des Langues (TAL) (Torres, 2014, 2011, Moreno et al 2020-23) ;
- Représentations classiques et denses (plongements des mots) venant de l'apprentissage profond (DL) (Martin et al., 2020 ; Avendano-Garrido, 2024).

Nous allons combiner et combiner la puissance des méthodes DL et l'explicabilité et simplicité des méthodes TAL et RI (Akani 2022).

Enfin une évaluation des modèles s'avère indispensable pour mesurer les résultats produits par des traitements informatiques efficaces.

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.

Les résultats du projet seront disséminés plus largement par le biais d'une application en ligne de concordancier, cqpweb (Hardie, 2021), dont une instance est hébergée sur les serveurs de l'université, et qui fonctionnent déjà pour la mise en ligne corpus mono- et bilingues, avec et sans annotations grammaticales et lexicales. Dans un premier temps, les résultats obtenus permettront la mise en ligne d'un corpus nahuatl monolingue, avec la possibilité d'y intégrer des métadonnées sous forme de balises xml, selon les spécificités textuelles des données. Dans un deuxième temps, il est envisagé de procéder à un alignement de corpus de nahuatl traduits, ce qui permettra la mise en ligne d'un corpus requêtable (traductions/source). L'application hébergée cqpweb permet un accès de l'extérieur, de manière sécurisée, ainsi que la création de comptes prévus pour un travail sur des corpus spécifiques. Au-delà de la génération de concordances, elle permet le calcul de fréquences, la distribution, les collocations, etc. selon un choix de méthodes statistiques. Cet outil pourra servir pour des travaux ultérieurs sur le projet, et la diffusion des résultats auprès de la communauté scientifique.

Également, une attention importante sera portée aux publications dans des congrès scientifiques nationales et internationales des résultats obtenus de l'ensemble du projet.

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, et l'Apprentissage automatique (en particulier apprentissage profond). La statistique

classique et la représentation au moyen des graphes seront de grande utilité dans cette étude.

6. Partenariats extérieurs envisagés

Nous avons établit des collaborations avec l'Universidad Veracruzana (Facultad de Matematicas et avec l'Instituto de Investigaciones en Educacion) afin de réaliser nos recherches conjointes sur le domaine. Nous allons poursuivre ces collaborations en vue de signer une convention de recherche entre nos 2 institutions.

7. Objectifs de pérennisation du projet

Ce projet s'inscrit déjà dans un appel précédent soutenu par l'Agorantic (NAHU). Cependant, ce projet présente plusieurs difficultés déjà évoquées ; d'où la nécessité de prolonger ce projet pour bien boucler les tâches de constitution du corpus et la modélisation des algorithmes. La demande d'une subvention interne à l'Universidad Veracruzana pour compléter les mobilités internationales est en cours. Également, une demande de bourse ministérielle de la thèse concernant l'étude et développement d'outils pour le nahuatl est actuellement en cours.

8. Expression des besoins en assistance informatique

Nous aurons besoin de 2 stagiaires (L o M) en informatique pour mettre en place les modèles ainsi que réaliser les tests et leurs évaluations quantitatives vis à vis des corpus de référence ou des annotations humaines. Également nous aurons besoin de visualisations adéquates permettant d'évaluer qualitativement les résultats.

9. Évaluation du projet

Informatique/TAL et Nahuatl :
- Luis MENESES LERÍN <meneseslerin.luis@gmail.com> (LINGUISTIQUE APPLIQUEE, Univ ARRAS France)
- Ximena Gutierrez, ximena.gutierrezv@gmail.com UNAM
- Karla AVILES <karla.j.aviles@gmail.com> (LINGUISTIQUE NAHUATL, INALCO Paris France)

10. Budget (€) prévisionnel *

	Brève description	Montant
Missions	Congrès CORIA et/ou TALN	3000 euros
Consommables, petits matériels**		
Organisation de réunions		
Stages*, **vacations**	Développement en python et Tests des outils informatiques	2 stages x 2 mois à 1200 = 2400 euros
Prestations de service	Annotation des corpus	2600
Budget total		8000
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

Mon directeur d'unité est informé du dépôt de ce projet x

Bibliographie

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

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

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

(4) J-M Torres-Moreno, M El-Bèze, F Béchet, N Camelin, Fusion probabiliste appliquée à la détection et classification d'opinions, DEFT'09, Paris, France, 22 juin 2009, 15p.

(5) Daniel Watson, Nasser Zalmout and Nizar Habash, Utilizing Character and Word Embeddings for Text Normalization with Sequence-to-Sequence Models, 2018 EMNLP pages 837–843

(6) L Moreno, J-M Torres-Moreno. LiSSS: a new multi-annotated multi-emotional corpus of Literary Spanish Sentences, CyS, 24(3) :1139–1147, 2023

(7) Hardie, A., CQPweb — Combining Power, Flexibility and Usability in a Corpus Analysis Tool. International Journal of Corpus Linguistics, vol. 17, nº 3, 2012, p. 380-409. CrossRef, https://doi.org/10.1075/ijcl.17.3.04har

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

(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) Avendano-Garrido, M-L, J-M Torres-Moreno, Ramirez J. Réduction de la dimensionnalité des embeddings au moyen des ensembles approximatifs. CORIA 2024. doi:10.24348/coria.2024.abstract_11

(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. 27th International Conference on Computational Linguistics (COLING'18), 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

(20) Figueroa-Saavedra, M. Amapowalistli iwan tlahkuilolewalistli. Tlamachtilamoxtli. Universidad Veracruzana, Xalapa 2024