APPEL À PROJETS 2026 – PROJETS *DONNÉES & OPEN DATA*

FÉDÉRATION DE RECHERCHE AGORANTIC

« Culture, Patrimoines, Sociétés Numériques »

1. **Ne pas dépasser 5 pages ;**
2. **Supprimer les indications en italique** lors du remplissage du formulaire ;
3. Transmettre le fichier au **format PDF** intitulé : ACRONYME-AAP2-Données-Agorantic-2026, où ACRONYME est à remplacer par l'acronyme du projet ;
4. Envoyer le fichier à fr-agorantic@univ-avignon.fr **avant le 15 octobre 2025** ;
5. **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.**

Titre	*Expansion de corpus textuels en nahuatl et leur modélisation via Grammaires non contextuelles et duplication de données pour le apprentissage des LLM.*		
Acronyme	***NAWA***		
Porteur/ porteuse	*Juan-Manuel Torres-Moreno*	juan-manuel.torres@univ-avignon.fr	
	Laboratoire Informatique d'Avignon		
Gestionnaire du laboratoire	**Michèle MANEN**	michele.manen@univ-avignon.fr	
Liste exhaustive des participant·es identifié·es	**Prénom & nom**	**Laboratoire**	**Courriel**
	1. Graham Ranger 2. Martha-Lorena Avendano-Garrido 3. Miguel Figueroa-Saavedra Ruiz 4. Rémy Kessler	1. Lab Identité Culturelle, Textes et Théatrâlité (ICTT) 2,3. Univ de Veracruz, Fac Mathématiques et Inst de Inv. en Educacion 4. LIA-Avignon	{Graham.Ranger,Remy.Kessler}@univ-avignon.fr {migfigueroa, maravendano}@uv.mx
Budget demandé	*8000 €*		
Résumé ***Max. 1 000 caractères, espaces compris***	Nous cherchons à développer des algorithmes pour l'expansion des corpus existants en nahuatl. En effet, les documents disponibles dans cette pi-langue (peu dotée de ressources), sont rares et la diversité linguistique est très grande. Également, il y a plusieurs graphies qui compliquent la tâche de collecte des données. Cependant, le projet NAHU² a permis de construire un système orthographique unifiée. Les graphies sont diverses car il y a au moins 4 régions (Centrale, Nord, Pacifique, Golfe du Mexique) où le nahuatl a trop évolué (les vocabulaires différent et leurs caractéristiques grammaticales sont distinctes). La langue est complexe au niveau des structures grammaticales, et les ressources informatisées sont pratiquement inexistantes. Nous voulons construire un système d'expansion artificielle des corpus, en utilisant des Grammaires Libres de Contexte (CFG) et la duplication contrôlée des données. Nous nous appuierons sur des techniques linguistques, TAL et d'IA profonde à ces propos.		

6. Contexte, positionnement, objectif(s)

La langue nahuatl (ou nawatl) est une des langues nationales du Mexique avec l'espagnol, et elle est parlée par environ 2,5 millions de personnes (depuis le Canada, EU, Mexique, jusqu'au le Nicaragua). Le nahuatl 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 auprès des publics. Pour cette raison, ce défi socio-linguis-tique, implique une opportunité pour développer d'applications informatiques intéressants pour le public nahuaphone natif ou non. Cette culture écrite nécessite des outils d'IA qui surmontent la diversité graphique et dialectale, ce qui permettra la génération de corpus textuels de taille adéquate, l'analyse d'informations et l'apprentissage automatique. 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 en qualité) pour l'apprentissage automatique. En effet, nos projets NAHU et NAHU^2 ont servit à la compréhension basique du nahuatl (suivit des cours en ligne, exercices, etc) (Figueroa et al 2014, 2020-24); au développement des algorithmes de compression des embeddings (Avendano et al, 2024) et à la constitution du corpus pi-yalli. Le nahuatl étant une langue agglutinante, nous utilisons des embeddings de mots et de tokens (Watson, 2018), afin de pouvoir extraire les caractéristiques sémantiques fondamentales (en particulier la racine des verbes et leurs conjugaisons).

Dans ce projet de continuité nous nous proposons d'utiliser des techniques d'IA et de linguistique générative pour constituer des corpus en nahuatl adéquats à l'apprentissage profond. Nous savons maintenant que les 6M de tokens du corpus pi-yalli, issu du projet NAHU^2 sont pour l'instant adéquats pour l'apprentissage des embeddings statiques et carrément insuffisants pour ceux contextuels. En effet, on a besoin d'entre 10M-100M de tokens pour commencer un apprentissage avec des transformateurs. Or, quels transformateurs sont les meilleurs ou les plus simples à adapter ou à utiliser dans le cas du nahuatl ? Quels tokeniseurs utiliser ? Les grammaires formelles (Context-Free Grammars, CFG) peuvent-elles aider à produire des textes adéquats ? Dupliquer les corpus n'a pas été tenté dans les pi-langues... (Lee et al 2022) peut-elle servir à mieux identifier les structures grammaticales prototypiques ? 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 la grammaire du nahuatl dans des CFG, mais elle n'existent pas : il faut les créer. Ensuite, les utiliser en mode génératif (avec ou sans duplication) pour produire des phrases ayant en plus de la grammaticalité, une sémantique acceptable pour nos propos. Nous pensons que les algorithmes développés dans le projet NAWA serviront à constituer des corpus de taille adéquate pour les nouveaux transformateurs que nous allons développer : le système BERTL qui sera crée par et pour le nahuatl.

7. Questionnement scientifique

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

i/ Création de grammaires non contextuelles pour le nahuatl : il faut introduire des grammaires modélisant les structures grammaticales nahuatl les plus utilisés : SVO (Sujet-Verbe-Objet). La problématique scientifique concerne l'agglutination et le polysynthetisme du nahuatl en plus de les diverses variantes dialectales.

ii/ Duplication des données: La problématique scientifique concerne la duplication contrôlée des données textuelles nahuatl et plus précisément, la production automatisée de données textuelles par incréments, thématiques, rhétorique et dialectale. Les données seront produites dans un contexte monolingue (nahuatl).

iii/ Créer des corpus textuels réalistes en nahuatl : avec un volume considérable, ayant comme objectif l'apprentissage automatique sur des corpus de volume adéquats.

Dans ce projet nous allons nous concentrer sur les verrous scientifiques i/, ii/ et iii/, qui concernent la création de corpus artificiels réalistes et élargis (FR-NAH).

8. Méthodologie

Nous utiliserons de outils de Traitement Automatique de Langues (TAL) et d'apprentissage profond que le LIA maîtrise depuis longtemps (Torres et al, 2009), afin de pouvoir constituer et élargir artificiellement des corpus en nahuatl. Ces outils sont facilement adaptables aux besoins des expériences proposées. Nous nous proposons de créer des grammaires non contextuelles (CFG) pour modéliser le nahuatl, langue agglutinante et polysynthétique dont la grammaticalité gravite autour du verbe. Nous allons également étudier plusieurs représentations denses (embeddings de mots, des caractères, etc.) ainsi que leur segmentation à base de tokenisations adaptées et spécialisées pour le nahuatl. La duplication contrôlée (à l'identique, par thématique, dialectale et rhétorique) sera aussi étudiée et appliquée à l'expansion des corpus. 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 :

- Traitement Automatique des Langues (TAL) (Torres, 2014, 2011, Moreno et al 2020-23) ;

- Représentations denses (plongements des mots) venant de l'apprentissage profond (DL) (Martin et al., 2020 ; Avendano-Garrido, 2024).

- CFG, qui seront adaptées au nahuatl

Nous allons combiner la puissance des méthodes DL et simplicité des méthodes TAL. Enfin une évaluation des modèles s'avère indispensable pour mesurer les résultats produits par des traitements informatiques efficaces.

9. Résultats attendus et caractère innovant de la recherche

D'abord nous voulons constituer des corpus artificiels bien plus grands que ceux « authentiques ». Ils seront de grande taille et ils seront utilisées pour apprentissage profond à base de modèles statiques et contextuels (LLM). Les CFG nahuatl n'ont jamais vu le jour à notre connaissance, et la duplication de corpus dans des pi-langues non plus. Il faut dire que la déduplication des données est plutôt la règle dans les tau-langues (Lee at al 2022), mais nous voulons montrer le contraire dans le cas spécifique du nahuatl. 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 de similarité sémantique (que sera approximée par une similarité lexicale et des techniques de sémantique distributionnelle) et des calculs des divergences de distribution des probabilités. L'évaluation qualitative se fera au moyen de lecture et d'analyses directes faites par des experts dans la langue nahuatl. Pour cela, nos collègues de l'Univ. Veracruzana (Inst de Investigacion en Educacion) nous apportent déjà un soutient logistique important 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 des corpus authentiques et artificiels (parallèles ou pas), avec et sans annotations POS et lexicales. Nos résultats obtenus permettront d'enrichir la mise en ligne de nos données nahuatl dans le site web : **https://demo-lia.univ-avignon.fr/pi-yalli**, avec la possibilité d'y intégrer des métadonnées xml et les embeddings résultants. Dans un deuxième temps, il est envisagé de procéder à une fusion de corpus authentiques et artificiels, ce qui permettra l'apprentissage de modèles statiques et LLM. Ces modèles pourront servir de base pour résoudre des tâches TAL définies ultérieurement dans le projet (similarité sémantique , analyse des sentiments, détection d'entités nommées et résumé automatique), et la diffusion des résultats pour la communauté scientifique. Une attention importante sera portée aux publications dans des congrès et colloques scientifiques nationales et internationales des résultats obtenus de l'ensemble du projet, comme cela a été toujours fait lors du projet NAHU^2.

10. Dimension interdisciplinaire

Notre projet s'inscrit fortement dans les axes suivants : A. Données et corpus : constitution, exploitation et valorisation ; B. Description, modélisation, théorisation. Il est axé sur le TAL, la linguistique computationnelle et les corpus, et l'Apprentissage profond. La statistique et les représentations au moyen des graphes seront de grande utilité dans cette étude.

11. Partenariats extérieurs envisagés

Nous avons déjà établit des collaborations avec l'Universidad Veracruzana (Facultad de Matematicas et avec l'Instituto de Investigaciones en Educacion). Mme AVENDANO et M FIGUEROA co-dirigent actuellement la thèse de M GUZMAN, portant sur l'analyse automatisée de nahuatl. Nous réalisons des recherches conjointes avec le Mexique sur le domaine et nous allons poursuivre et approfondir ces études.

12. Objectifs de pérennisation du projet

Le projet NAWA s'inscrit déjà dans un appel précédent soutenu par l'Agorantic (NAHU^2). Cependant, ce projet présente plusieurs difficultés spécifiques ; d'où la nécessité de prolonger le projet NAHU^2 pour bien boucler les tâches d'expansion des corpus et l'amélioration des algorithmes. Également, une demande de bourse ministérielle pour la poursuite de la thèse actuelle, qui concernera la traduction automatisée nahuatl-français est

actuellement en préparation.

13. Expression des besoins en assistance informatique

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

14. Expert·es suggéré·es pour l'évaluation du projet

- Karla AVILES <karla.j.aviles@gmail.com> (LINGUISTIQUE NAHUATL, INALCO Paris France)
- Luis MENESES LERÍN <meneseslerin.luis@gmail.com> (LINGUISTIQUE APPLIQUEE, Univ ARRAS France)

15. Budget (€) prévisionnel *

	Brève description	Montant
Missions	Congrès CORIA'26, TALN'26, MICAI'26	3000
Consommables, petits matériels**		
Organisation de réunions		
Stages*, vacations	Développement des outils informatiques et vérification des tests	2 stagesx2 mois 1200=2400
Prestations de service	Annotation des corpus	2600
Budget total		8000
Co financements le cas échéant		
Budget demandé à Agorantic		**8000**
Recettes extérieures	**Mission UV 2026**	**2000**

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

Bibliographie

(1) Abdillahi, N., Nocera, P., and Torres, J. M. (2006). Boites a outils TAL pour les langues peu informatisées : Le cas du Somali. In Journées d'Analyses des Données Textuelles, Besançon, France.

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

(3) Flores Nájera, L. (2019). La gramática de la clausula simple en el náhuatl de Tlaxcala. PhD thesis, CIESAS.

(4) 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

(5) 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.

(6) Figueroa, M., Alarcón, D., Bernal, D. et Hernández, J.A. (2014). The incorporation of national indigenous languages into the academic development of universities: The experience of the UV. Rev. Educación Superior 43(171), 67-92

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

(8) Guzmán, J.-J., Torres, J.-M., Ranger, G., Garrido, M. L., Figueroa, M., Quintana, L., González, C.-E., Linhares, E., Velázquez, P., and Moreno, L.- G. (2025b). π-yalli: un nouveau corpus pour le nahuatl / Yankuik nawatlahtolkorpus pampa tlahtolmachiotl. In TALN Marseille, pages 802–816. ATALA.

(9) Guzmán, J.-J., Vázquez, J., Torres, J.-M., Ranger, G., Garrido, M. L., Figueroa, M., Quintana, L., Velázquez, P., and Sierra, G. (2025c). A symbolic algorithm for the unification of nawatl word spellings. In MICAI'25, page 12p. SMIA.

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

(11) 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

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

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

(14) Lastra de Suárez, Y. (1986). Las áreas dialectales del náhuatl moderno. UNAM, Instituto de Investigaciones Antropológicas, Mexico.

(15) Launey, M. (1978). Introduction à la langue et à la littérature aztèques, volume 1. L'Harmattan, Paris.

(16) Lee K. et al. 2022. Deduplicating Training Data Makes Language Models Better. In Proceedings of the 60th Annual Meeting of the ACL (Volume 1: Long Papers), pages 8424–8445, Dublin, Ireland.

(17) Liu, C., Zhang, H., Zhao, K., Ju, X., and Yang, L. (2024). LLMEmbed: Rethinking lightweight LLM's genuine function in text classification. 62nd Annual Meeting ACL (Volume 1: Long Papers), pages 7994–8004, Bangkok.

(18) 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

(19) Micheli, V., d'Hoffschmidt, M., and Fleuret, F. (2020). On the importance of pre-training data volume for compact language models. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, EMNLP, 7853–7858, Online. ACL.

(20) Mager, M., Gutiérrez, X., Sierra, G. et Meza, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. 27th COLING'18, pp. 55–69, Santa Fe, New Mexico, USA. ACL.

(21) Olko, J. and Sullivan, J. (2016). Bridging gaps and empowering speakers: An inclusive, partnership-based approach to nahuatl research and revitalization. Integral strategies for language revitalization, pages 347–386.

(22) Pugh, R., Wing, C., Juárez, M. X., Márquez, Á., and Tyers, F. (2025). Ihquin tlahtouah in tetelahtzincocah: An annotated, multi-purpose audio and text corpus of western sierra Puebla Nahuatl. NAACL: Human Language Technologies (Volume 1: Long Papers), 3549–3562, Albuquerque, New Mexico. ACL.

(23) Smith, N. A. and Johnson, M. (2007). Weighted and probabilistic context-free grammars are equally expressive. Computational Linguistics, 33(4):477–491.

(24) Torres, J.-M., Avendaño, M.-L., Figueroa, M., Ranger, G., González, C., Linhares, E., Velazquez, P., Quintana, L., and Guzmán, J.-J. (2024a). NAHU^2: Un nouveau corpus pour le Nahuatl. 18èmes J Inf. Région Centre-Val de Loire

(25) Torres-Moreno, JM, Automatic Text Sumarization, London Wiley, 2014

(26) Zimmermann, K. (2019). Estandarización y revitalización de lenguas amerindias: funciones comunicativas e ideológicas, expectativas ilusorias y condiciones de la aceptación. Revista de Llengua i Dret, J of Language and Law, 71:111–122.