NAHU^2: Un nouveau corpus pour le Nahuatl

Juan-Manuel Torres-Moreno, Martha-Lorena Avendaño-Garrido, Miguel Figueroa-Saavedra, Graham Ranger, Carlos-Emiliano González-Gallardo, Elvys Linhares Pontes, Patricia Velázquez-Morales, Ligia Quintana Torres, Juan-José Guzmán-Landa

NAHU²: Un nouveau corpus pour le Nahuatl

Juan-Manuel Torres-Moreno[1] Martha-Lorena Avendaño-Garrido[2] Miguel Figueroa-Saavedra[2] Graham Ranger[3] Carlos-Emiliano González-Gallardo[4] Elvys Linhares Pontes[5] Patricia Velázquez-Morales Ligia Quintana Torres[2,1] Juan-José Guzmán-Landa[1]

(1) LIA, Université d'Avignon, Avignon, France
(2) Universidad Veracruzana, Veracruz, Mexique
(3) ICTT, Université d'Avignon, Avignon, France
(4) LIFAT/CESR, Université de Tours, Tours, France
(5) Trading Central Labs, Trading Central, Paris, France
{juan-manuel.torres,graham.ranger}@univ-avignon.fr, murdoocc7@gmail.com
{maravendano,migfigueroa,liquintana}@uv.mx
gonzalezgallardo@univ-tours.fr elvys.linharespontes@tradingcentral.com
patricia_velazquez@yahoo.com

1 Introduction

Le projet NAHU² (collaboration Franco-Mexicaine) vise à constituer un corpus approprié d'apprentissage et de tests qui permettra par la suite le développement des ressources informatiques pour la langue Nahuatl. En effet, le Nahuatl est une langue peu dotée de ressources informatiques malgré le fait qu'elle soit une langue vivante et parlée pour environ 2 millions de personnes.

Nous avons décidé de constituer un corpus qui permettra effectuer des recherches sur le Nahuatl en vue de construire des modèles de langue (contextualisés et dynamiques ou pas) qui permettront à leur tour le développement des outils telles que :

— un unificateur de graphies ;
— un segmenteur de mots ;
— un analyseur grammaticale POS
— un résumeur basé de texte sur le contenu (Torres-Moreno, 2014).

Éventuellement si la taille du corpus le permet, d'envisager e) un traducteur (avec ou sans apprentissage) Nahuatl-Français.

1.1 Le Nahuatl

Il s'agit d'une langue autochtone de la famille uto-nahua, parlée par des millions de personnes au Mexique et dans d'autres régions d'Amérique centrale. C'est une langue parlée au Mésoamérique depuis le Vè siècle, étant la langue nationale la plus parlée, après l'espagnol, au Mexique avec 1.651,958 nahuaparlants (INEGI, 2020).

La figure 1 montre les variantes linguistiques du Nahuatl au Mexique. Pourtant, aujourd'hui c'est

FIGURE 1 – Variantes linguistiques du Nahuatl au Mexique

une langue considérée, selon sa variante, vulnérable ou en danger de disparition (UNESCO, 2012). Et ce, malgré les efforts que les communautés Nahuas ont déployé pour que leur langue soit utilisé à l'oral et à l'écrit dans l'industrie éditoriale, l'enseignement supérieur, les médias et les réseaux sociaux (Aguilar Santiago & García Zúñiga, 2023; Farfán, 2011; Saavedra & Martínez, 2023; Olko & Sullivan, 2014). Nonobstant son riche héritage, le Nahuatl fait face à des défis importants en raison de son statut de langue minoritaire et de la pénurie de ressources informatiques disponibles pour sa préservation et sa diffusion. Le Nahuatl est donc une langue-π ou langue peu dotée de ressources informatiques.

2 Corpus NAHU²

Nous avons collecté un ensemble de documents venant de plusieurs sources (i.e., PDF, fichiers texte, Microsoft Word, sites web et wiki) et des codifications hétérogènes (i.e., utf8 et utf16), qui a posé des défis informatiques. La structure des documents étant variée, nous les avons traité semi-automatiquement afin d'éliminer des entêtes, indices, tables, références et des paragraphes dans des langues autres que le Nahuatl. Nous avons aussi inclus les textes Nahuatl du corpus Axolotl (Gutierrez-Vasques *et al.*, 2016). Nous avons obtenu un corpus d'environ 1,5 millions de *tokens* et 12 millions de caractères. Les variétés incluses correspondent principalement aux variétés parlées dans l'État de Veracruz (Nahuatl centrale, Nahuatl de La Huasteca) également partagé avec d'autres États du centre et du nord du pays ; et de façon moins importante à la variété Nahuat du sud de Veracruz et de Puebla, et à la variété *tecpillahtolli* –un registre cultivé– utilisées entre le XVIe et le XIXe siècles employé dans les textes imprimés. Pour cette raison, des textes avec différents alphabets utilisés aujourd'hui et dans le passé sont inclus. Une fois constitué, sous forme de texte brut avec métadonnées identifiant chaque texte, le corpus sera mis en ligne pour une consultation par mots, suites de mots, regex, etc., par le biais de l'application cqpweb, interface graphique pour le Corpus Query Processeur (Evert & Hardie, 2011). Dans un deuxième temps, il est prévu d'indexer une deuxième version du corpus, enrichie d'annotations grammaticales et de nouvelles métadonnées, dès que les outils seront disponibles.

3 Modèles

Les représentations vectorielles denses capturent les relations sémantiques entre les mots de manière sophistiquée et précise, permettant une compréhension du langage plus nuancée et efficace (Almeida & Xexéo, 2023). Ces représentations sont essentielles pour des applications nécessitant une compréhension sémantique avancée, telles que la reconnaissance des entités nommées, l'analyse du sentiment (Linhares Pontes *et al.*, 2018) et la classification de textes. Dans notre étude, nous nous concentrerons initialement sur les modèles Word2Vec (Mikolov *et al.*, 2013) et FastText (Bojanowski *et al.*, 2017). Ensuite, dans un deuxième temps, nous nous pencherons sur les modèles de langue contextualisés tels que BERT (Devlin *et al.*, 2019). Chaque modèle ayant ses spécificités, leurs performances varient selon la langue, le domaine, les nuances sémantiques et la taille du corpus étudié. Nous évaluerons donc ces modèles pour le Nahuatl afin de mesurer leur capacité à produire des représentations précises et cohérentes.

4 Protocole d'évaluation

Nous avons proposé un protocole d'évaluation de similitude sémantique. Étant donné les 25 termes de référence du tableau 1, chacun ayant associé une liste de 5 termes candidats, il a été demandé à 30 nahuaparlants de trier sémantiquement les listes de candidats du plus proche au plus éloigné. Chaque candidat a reçu une note de 1 à 5 (1 jugé le plus proche sémantiquement à la référence et 5 le plus éloigné). Ceci a permis de créer un ensemble de rangs.

itatzin son père	**nemi** on y habite	**nepa** là-bas/ça	**miki** mourir	**acontle** jarre	**onikkak** j'ai entendu	**tikitta** tu le vois	**noyollo** mon cœur	**tototl** oiseau
wewetkeh vieux/vieille	**miyak** beaucoup	**mawistik** étonnant	**altepetl** ville	**ilhuicac** au ciel	**piyalli** bonjour	**melawak** je vais bien	**tekitl** travail	**noyolikni** mon ami
axcan maintenant	**onipeh** j'ai commencé	**tzopelik** sucré	**tamalli** tamal	**amatl** papier	**istak** blanc	**tlakentli** linge		

TABLE 1 – Liste de 25 de référence du corpus

Afin d'évaluer l'accord entre les annotateurs, le coefficient de concordance W de Kendall a été utilisé. Ce coefficient permet d'évaluer la cohérence entre plus de 2 classements, et représente une extension du coefficient de corrélation τ de Kendall, conçu spécifiquement pour mesurer le degré de concordance entre 2 classements. Une valeur de 1 indique une concordance parfaite (tous les classements sont identiques), et une valeur de 0 indique une absence totale de concordance, ce qui signifie que les positions dans les classements sont complètement incohérentes.

Les résultats (figure 2) montrent que le W de Kendall pour les 25 références varie de 0,08 pour *tamalli* (tamal), qui a obtenu la valeur la plus basse (probablement en raison de l'inclusion d'options de termes candidats trop locaux ou méconnaissables pour la plupart des locuteurs), à 0,52 pour la référence *onipeh* (j'ai commencé), avec la valeur la plus élevée. La moyenne W pour tous les candidats est d'environ 0,26.

Les résultats des modèles Word2Vec et FastText sont encore en cours, et nous devons raffiner le protocole présenté afin de focaliser sur les couples références-candidats les mieux adaptés afin de bien paramétrer les algorithmes. Nous proposerons également d'autres tâches TAL pour évaluer la

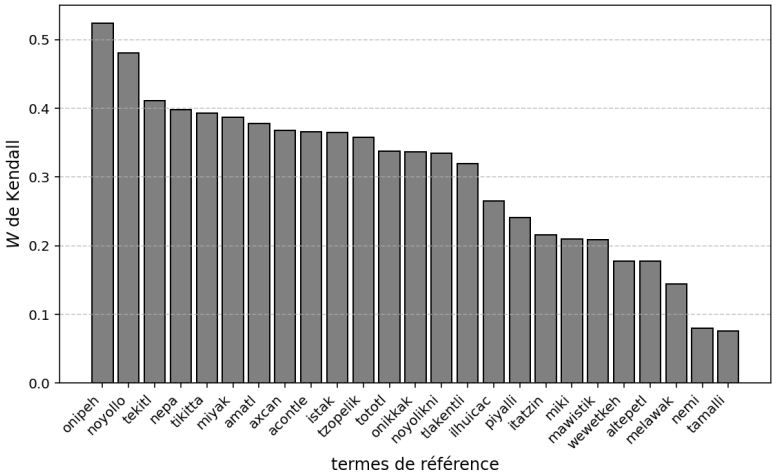

FIGURE 2 – W de Kendall mesurant l'accord entre les 30 nahuaparlants par mot de référence

qualité du corpus. Notamment nous utiliserons des modèles type BERT pour établir des étalons plus adéquats d'évaluation.

5 Conclusion

Bien que le corpus NAHU² soit encore en développement (nous sommes en train de dupliquer son volume), nous pensons qu'il est une ressource intéressante pour étudier l'impact de la taille d'un corpus dans l'apprentissage de modèles de langue du Nahuatl. Il nous permettra de développer des outils d'analyse classiques et probablement des modèles IA légers que nous diffuserons auprès de la communauté scientifique. En outre, l'utilisation naissante et croissante du Nahuatl dans les réseaux sociaux, l'industrie de l'édition, les programmes éducatifs universitaires et la diffusion scientifique rendent ces outils de plus en plus nécessaires pour accéder et gérer l'information disponible sur Internet.

Cette accessibilité permettra de relier différentes communautés de locuteurs situés dans différentes régions et pays, en plus de permettre la circulation des connaissances exprimées dans cette langue auprès des étudiants et des spécialistes. Cela représente une puissante impulsion pour l'activation et la mise à jour de cette importante langue minoritaire.

Remerciements

Ce travail a été soutenu par les projets NAHU et NAHU² financés par FR Agorantic et Intermedius d'Avignon Université.

Références

AGUILAR SANTIAGO C. A. & GARCÍA ZÚÑIGA H. A. (2023). Tecnologías del lenguaje aplicadas al procesamiento de lenguas indígenas en méxico : Una visión general. *Lingüística y Literatura*, (84), 79–102.

ALMEIDA F. & XEXÉO G. (2023). Word embeddings : A survey.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EVERT S. & HARDIE A. (2011). Twenty-first century corpus workbench : Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, p. 1–21 : Citeseer.

FARFÁN J. A. F. (2011). El proyecto de revitalización, mantenimiento y desarrollo lingüístico y cultural : resultados y desafíos. In *Indigenous Languages of Latin America Actas del Primer Simposio sobre Enseñanza de Lenguas Indígenas de América Latina*, p. 114.

GUTIERREZ-VASQUES X., SIERRA G. & POMPA I. H. (2016). Axolotl : a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 4210–4214.

INEGI (2020). Censo de poblaciÓn y vivienda 2020.

LINHARES PONTES E., HUET S., LINHARES A. C. & TORRES-MORENO J.-M. (2018). Predicting the semantic textual similarity with Siamese CNN and LSTM. In P. SÉBILLOT & V. CLAVEAU, Éds., *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, p. 311–320, Rennes, France : ATALA.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, p. 3111–3119, Red Hook, NY, USA : Curran Associates Inc.

OLKO J. & SULLIVAN J. (2014). Toward a comprehensive model for nahuatl language research and revitalization. In *Annual Meeting of the Berkeley Linguistics Society*, p. 369–397.

SAAVEDRA M. F. & MARTÍNEZ J. Á. H. (2023). *In nawatlahtolli ipan interkoltoral tlamachtilistli itech Veracruz : owihkayotl iwan chikawakayotl*. Universidad Veracruzana.

TORRES-MORENO J.-M. (2014). *Automatic text summarization*. John Wiley & Sons.

UNESCO P. (2012). Atlas des langues en danger dans le monde.