π-YALLI : un nouveau corpus pour des modèles de langue nahuatl / *Yankuik nawatlahtolkorpus pampa tlahtolmachiotl*

Juan-José Guzmán-Landa[1] Juan-Manuel Torres-Moreno[1]
Martha Lorena Avendaño Garrido[2] Miguel Figueroa-Saavedra[2]
Ligia Quintana-Torres[2,1] Graham Ranger[1] Carlos-Emiliano González-Gallardo[3,1]
Elvys Linhares-Pontes[4] Patricia Velázquez-Morales[5] Luis-Gil Moreno-Jiménez[6]
(1) Laboratoire Informatique d'Avignon & ICTT, Université d'Avignon (France)
{juan-manuel.torres,juan-jose.guzman-landa,graham.ranger}@univ-avignon.fr
(2) Fac. de Matemáticas & IIE, U Veracruzana (Mexique), {maravendano,migfigueroa,liquintana}@uv.mx
(3) LIFAT, Université François Rabelais à Tours (France), gonzalezgallardo@univ-tours.fr
(4) Trading Central Labs, Trading Central (France), elvys.linharespontes@tradingcentral.com
(5) patricia_velazquez@yahoo.com (6) STIH luis-gil.moreno_jimenez@sorbonne-universite.fr

RÉSUMÉ

Le nahuatl ou nawatl, dispose de peu de ressources informatiques, bien qu'il soit une langue vivante parlée par environ deux millions de personnes. Nous avons construit π-YALLI, corpus qui permet de mener des recherches et de développer des modèles de langue (ML) dynamiques et statiques. Nous avons mesuré la perplexité de π-YALLI, évalué la performance des ML les plus récents comparés aux résultats d'un corpus de similitude sémantique annoté manuellement. Les résultats montrent la difficulté de travailler sur cette π-langue, tout en ouvrant des perspectives intéressantes pour l'étude d'autres tâches de Traitement Automatique des Langues (TAL) portant sur le nahuatl.

ABSTRACT

π-YALLI : a new corpus for Nahuatl Language Models

The Nahuatl is a language with few computational resources, despite the fact that it is a living language spoken by around two million people. We built π-YALLI, a corpus that enables research and development of dynamic and static Language Models (LM). We measured the perplexity of π-YALLI, evaluating state-of-the-art LM performance on a manually annotated semantic similarity corpus relative to annotator agreement. The results show the difficulty of working with this π-language, but at the same time open up interesting perspectives for the study of other NLP tasks on Nahuatl.

MOTS-CLÉS : Nahuatl ; Similarité sémantique ; Accord entre annotateurs ; ML ; π-langues.

KEYWORDS: Nahuatl ; Semantic similarity ; Annotators' agreement ; LM ; π-languages.

1 Introduction

Le **nahuatl**, **nawatl** ou **mexicano** (**nawatlahtolli** en nahuatl), langue autochtone de la famille Uto-Nahua (Smith, 2002; de Durand-Forest *et al.*, 1995) est parlé par un grand nombre de personnes au Mexique et d'autres régions d'Amérique [1]. Langue parlée au Mésoamérique depuis le Vème siècle, elle est la langue nationale la plus parlée au Mexique, après l'espagnol, avec 1 651 958 nahuaphones

1. https://fr.wikipedia.org/wiki/nahuatl

(INEGI, 2020) et plus de 2,5 millions de personnes dans la nahuaphonie. D'après INALI [2], il existe 30 variétés linguistiques de nahuatl parlées au Mexique. De nos jours, certaines variétés de nahuatl (Moseley, Christopher (ed.) & Nicolas Alexandre (cart.), 2012) sont en danger de disparition. Ceci malgré les efforts continus que les communautés Nahua ont déployés depuis 2003 – année de la reconnaissance du nahuatl comme langue nationale – pour que leur langue soit utilisée à l'oral et à l'écrit, dans l'industrie éditoriale, dans l'enseignement supérieur, les médias et les réseaux sociaux (Aguilar Santiago & García Zúñiga, 2023; Farfán, 2011; Figueroa-Saavedra & Hernández-Martínez, 2023; Olko & Sullivan, 2014). Le nahuatl étant une langue polysynthétique et agglutinante, les mots sont composés d'une racine verbale ou nominale et de morphèmes divers permettant de construire une signification. Un exemple d'agglutination est [3] :

Kuawtochtontli
kuaw-toch-ton-tli
bois-lapin-DIM-ABS
Petit lapin sauvage

Les relations syntaxiques entre mots sont établies par le biais de la valence du verbe et des connecteurs (particules). Ainsi, par exemple, on peut écrire :

Nikitta inin kalli tlen tikitta
Ni-k-itta inin kal-li tlen ti-k-itta
1SG.SUJ-3SG.OBJ-voir DEM maison-ABS que 2SG.SUJ-3SG.OBJ-voir
Je vois cette maison que tu vois

En utilisant une proposition subordonnée relative, ou bien écrire :

Nikitta tinechitta
Ni-k-itta ti-nech-itta
1SG.SUJ-3SG.OBJ-voir 2SG.SUJ-1SG.OBJ-voir
Je vois que tu me regardes

Ici la proposition subordonnée est l'objet du prédicat de la proposition principale. Ceux-ci peuvent également être composés par des regroupements établissant des nuances de sens, en plus des marqueurs et connecteurs discursifs (Figueroa-Saavedra, 2023). Par exemple [4] :

Ye onikittak amo otikpiyaya tlakualli
Ye o-ni-k-itta-k amo o-ti-k-piya-ya
déjà AOR-1SG.SUJ-3SG.OBJ-voir-PRFT NEG AOR-2SG.SUJ-3SG.OBJ-avoir-IMPF
*J'ai **déjà** vu que tu n'avais rien à manger*

niman axkan timayana,
niman axkan ti-mayana
alors maintenant 2SG.SUJ-avoir faim
***alors** aujourd'hui tu as faim*

wan moneki ma nimitzmaka tlaxkalli,
wan ϕ-mo-neki ma ni-mitz-maka tlaxkal-li
et 3SG.SUJ-REFL-vouloir INJ 1SG.SUJ-2SG.OBJ-donner tortilla-ABS
***et** tu as besoin que je te donne des tortillas.*

2. https://www.inali.gob.mx/clin-inali/html/l_nahuatl.html
3. Les gloses de Leipzig utilisées sont les suivantes : ABS : absolutif, AOR : aoriste, CAUS : causatif, DIR : directionnel, IMPF : imperfectif, INJ : interjection, NEG : négatif, OBJ : objet, DEM : démonstratif, PL : pluriel, PRFT : perfectif, REFL : réfléchi, SG : singulier, SUJ : sujet.
4. Marqueurs et connecteurs en gras en nahuatl et français.

Certains de ces mots sont appelés mots-phrases, puisque leur morphologie inclut le sujet et le prédicat, en plus d'informations sur les actants, et d'éléments modaux, relationnels et directionnels. Un exemple de polysynthèse est :

Axkan timitzoncochtiah
axcan ti-mitz-on-coch-tia-h
maintenant 1PL.SUJ-2SG.OBJ-DIR-dormir-CAUS-PL
Maintenant nous te faisons dormir loin.

Malgré son riche héritage, de nos jours le nahuatl fait face à des défis importants en raison de son statut de langue minoritaire et de la carence de ressources informatiques disponibles pour sa préservation et sa diffusion. Au lieu de faire référence au nahuatl comme une *langue minoritaire*, ce qui pourrait donner lieu à des biais négatifs, nous préférons considérer le nahuatl une π-langue, ou langue *peu dotée* de ressources informatiques (Berment, 2004; Abdillahi *et al.*, 2006) [5].

Le projet NAHU[2] (Torres-Moreno *et al.*, 2024a) vise à construire π-YALLI, un corpus adapté à l'apprentissage automatique et qui permettra le développement de ressources informatiques pour la langue nahuatl (Torres-Moreno *et al.*, 2024b). Nous avons introduit également un jeu de données pilote pour établir un protocole d'évaluation. À cette fin, une tâche de similitude sémantique de mots a été implémentée. Ce jeu de données pilote, annoté par plusieurs nahuaphones de différentes régions linguistiques, a servi à jauger la diversité des graphies possibles et la difficulté de traiter simultanément plusieurs variétés de nahuatl. C'est pourquoi nous avons évalué très précisément l'accord entre annotateurs. Ce corpus pilote a servi aussi pour juger la qualité des modèles de langue (ML) ayant appris sur π-YALLI. Nous avons ainsi constaté que les modèles statiques nahuatl sont encore compétitifs vis-à-vis des grands modèles de langue (GML) les plus performants.

Cet article est structuré de la manière suivante : en section 2 nous présentons un nouveau corpus nahuatl conçu pour l'apprentissage automatique (profond ou pas). En section 3 nous présentons le corpus pilote de nahuatl et le protocole d'évaluation des algorithmes les plus récents (GML vs. ML statiques). En section 4 nous montrons nos résultats, avant de conclure en section 5.

2 Le corpus π-YALLI

On relève quelques efforts antérieurs pour développer des corpus en nahuatl. Par exemple, le corpus *Axolotl* (Gutierrez-Vasques *et al.*, 2016) [6] est un corpus parallèle nahuatl-espagnol, portant sur deux variétés de nahuatl. Cependant, des facteurs tels que le caractère oral de cette langue, le manque de standardisation des graphies ou encore le nombre important de variétés, font que le nombre de ressources dont on dispose reste limité. Pour pallier ce manque, nous avons créé le nouveau corpus nahuatl π-YALLI (*bonjour !* en français) en collectant un ensemble de documents issus de différents sources, formats (pdf, txt, doc/odt, html, wiki) et encodages (iso-latin, us-ascii, utf8/16). La structure hétérogène des documents impose un traitement semi-automatique afin d'éliminer des fragments non informatifs. Ainsi, les en-têtes, les indices, les tables, plusieurs références bibliographiques et des paragraphes écrits dans des langues autres que le nahuatl ont été supprimés des documents. Le corpus π-YALLI [7] est constitué d'environ 394K phrases ; 6,12M *tokens* (chaînes de caractères séparées par

5. Par opposition aux τ-langues, *très bien* dotées de ressources informatiques et aux μ-langues *moyennement* dotées de ressources informatiques.
6. https://axolotl-corpus.mx
7. π-YALLI est téléchargeable à l'adresse https://demo-lia.univ-avignon.fr/pi-yalli

des espaces) et 48,5M de caractères plain texte, codé en utf8 (code ISO 639-3 nah) [8]. À différence des langues où les mots peuvent être séparés par des blancs, et en raison de sa nature polysynthétique, il est difficile de mesurer le nombre exact de mots nahuatl. En effet, un « mot » nahuatl est souvent constitué de plusieurs mots agglutinés. Par exemple **Koakalko**, compté comme 1 « mot » nahuatl, est en réalité construit avec **Koatl+kalli+ko** → **Koa(-tl)+kal(-li)+ko**, ou encore traduit par « dans la maison du serpent » (5 mots en français).

Les variétés incluses dans le corpus π-YALLI correspondent principalement à celles parlées dans l'État de Veracruz (nahuatl de la zone centrale et 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 nawat du sud de Veracruz et de l'État de Puebla, et à la variété *tecpillahtolli* – registre savant ou nahuatl classique – utilisée entre le XVIème et XIXème siècles et qui a été employé dans les textes imprimés. Des textes avec différentes graphies utilisées aujourd'hui et dans le passé ont été inclus dans le corpus. Nous avons unifié automatiquement ces graphies afin de limiter l'impact des variations dans l'apprentissage et la performance des modèles [9]. Les documents ont été classés empiriquement dans 16 catégories : documents historiques (HIS) ; Wikipédia (WIK) [10] ; littérature (LIT) ; poésie (POE) ; musique (MUS) ; documents légaux (LEG) ; linguistique (LIN) ; politique (POL) ; médecine (MED) ; économie (ECO) ; mémoires de master, travaux scolaires (EDU), religion (REL) ; cosmovision (COS) ; agriculture (AGR) ; phrases sans contexte (PHR) et documents scientifiques/techniques (TEC). Nous avons aussi inclus les nombreux paragraphes nahuatl appartenant aux versions disponibles du corpus Axolotl (Gutierrez-Vasques *et al.*, 2016) [11]. Quelques statistiques descriptives du corpus sont présentées dans la Table 1.

Catégorie	Nb doc	Variétés de nahuatl	Tokens	% Corpus
AGR	3	cen(2),hua(1)	7 828	0,13 %
COS	3	cen(1),hua(1),cla(1)	40 983	0,67 %
ECO	1	cen(1)	16 777	0,27 %
EDU	67	cen(66),hua(1)	276 763	4,52 %
HIS	54	cla(47),cen(6),pue(3)	645 406	10,54 %
LEG	20	cen(8),cla(3),pue(1),hid(2),hua(4)	341 945	5,58 %
LIN	9	cen(6),hua(2),cla(1)	368 511	6,02 %
LIT	52	cen(31),pue(10),gue(4)	881 369	14,39 %
MED	4	cen(2),hua(2)	14 248	0,23 %
MUS	5	cen(5)	4 306	0,07 %
PHR	49	cen(42), mix(7)	9 259	0,15 %
POE	11	cen(9), mix(2)	5 647	0,09 %
POL	2	mor(1),cen(1)	1 082	0,02 %
REL	29	cla(14),cen(4),pue(5),gue(1),oax(3),hua(2)	3 308 363	54,05 %
TEC	1	cen(1)	518	0,01 %
WIK	4 298	mélange de variétés	194 292	3,17 %
TOTAL	**4 608**	cen(185),cla(66),pue(19),hua(13),hid(2),oax(3),mor(1)	**≈6 121 000**	**100,0 %**

TABLE 1 – Statistiques sur les documents du corpus π-YALLI. cen=centrale, cla=classique, pue=Puebla, gue=Guerrero, oax=Oaxaca, mor=Morelos, hua=Huasteca, hid=Hidalgo ; entre () le nombre de documents par variété. Dans plusieurs cas il peut y avoir un mélange (mix).

8. https://iso639-3.sil.org/code/nah

9. La graphie utilisée a été implémentée avec des regexp en perl 5.0 et elle sert uniquement dans nos traitements informatiques. Aucune graphie ne fait encore consensus parmi les locuteurs.

10. https://nah.wikipedia.org, environ 4,3K petits articles.

11. Projets py-elotl, https://pypi.org/project/elotl et https://huggingface.co/datasets/somosnlp-hackathon-2022/Axolotl-Spanish-Nahuatl

Nous avons estimé la perplexité $P = 2^H$, entropie = H (Manning & Schütze, 1999) en unigrammes du corpus π-YALLI par rapport à d'autres corpus [12] de langues très bien dotées de ressources informatisées (voir Table 2). La perplexité a été calculée de deux façons : avec le corpus brut et avec le corpus après une normalisation (de la casse, balises, ponctuation, etc.) et l'unification de graphies. La perplexité élevée du nahuatl met en évidence une difficulté supplémentaire à travailler sur cette langue.

	nahuatl	nahuatl normalisé	anglais	français	portugais	espagnol
Entropie	13,46	**12,13**	11,93	11,42	11,86	11,92
Perplexité	11 236	**4 471**	3 908	2 745	3 719	3 872

TABLE 2 – Perplexité du nahuatl par rapport à quelques τ-langues.

3 Protocole d'évaluation sémantique

Notre étude cible plusieurs aspects liés aux documents nahuatl. Nous cherchons à établir une tâche étalon pour le corpus π-YALLI, à mesurer l'accord des annotateurs parlant différentes variétés, à estimer l'impact des graphies sur l'apprentissage des modèles et également à établir une corrélation entre la taille du corpus et la performance des ML, entre autres. À cette fin, un protocole d'évaluation permettant de mesurer la performance des algorithmes sur une tâche précise de similitude sémantique a été établi. Il consiste en trois étapes : 1/ création d'un corpus de similitude sémantique, annoté manuellement ; 2/ mesure de l'accord entre les annotateurs ayant participé à sa création, et 3/ mesure des performances des ML les plus récents sur ce corpus pilote.

3.1 Corpus pilote de similitude sémantique

Nous allons décrire ici la construction du corpus pilote de similitude sémantique de mots. Il sera fondamental pour mesurer la performance des modèles dans une tâche précise. Étant donné 23 termes de référence, chacun ayant associé une liste de 5 termes candidats, il a été demandé à 27 annotateurs nahuaphones de trier sémantiquement cette liste, du candidat le plus proche au plus éloigné de la référence (Torres-Moreno *et al.*, 2024a). Il faut préciser que nous avons retenu un nombre impair d'annotateurs pour éviter des potentiels classements ex-æquo. Chaque candidat a reçu une note de 1 à 5 (1 = terme jugé le plus proche sémantiquement à la référence, 5 le plus éloigné). Ceci crée un ensemble de 23×27=621 rangs, de 5 mots chacun. La sélection de 133 termes uniques [13] (23 références + 23×5 candidats - 5 doublons) a été réalisée selon différents critères :

Mots d'usage courant Ils sont exprimés dans trois catégories grammaticales : substantifs, verbes et particules, comprenant des noms d'ustensiles, d'aliments, de vêtements, de couleurs, de goûts, de qualités, de termes de parenté et de parties du corps.

Actions quotidiennes Elles sont définies au moyen des verbes transitifs, intransitifs, d'état et de mouvement, fléchis en numéro et formes verbales, et des particules adverbiales de nature quantitative et locative spatio-temporelle.

Expressions de salutation. Elles sont exprimées à l'aide des mots-phrases courants.

12. Français, espagnol et portugais de https://www.ortolang.fr/market/corpora/megalite, et anglais de https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2539

13. Il y a 5 mots répétés : (komitl, noyollo, tototl, tlapowalli, tekitl).

Ces termes ont été exprimés dans différentes variétés dialectales, y compris des formes caractéristiques du nahuatl central, de La Huasteca et du Sud, dans des formes caractéristiques du nahuatl savant ou littéraire (*tecpillahtolli*), et en utilisant différentes conventions alphabétiques employés par les locuteurs, mais avec une majorité de formes caractéristiques du nahuatl central, écrites avec un alphabet modernisé. L'expérience a été mise en place via une enquête en ligne, où les données fournies par les annotateurs ont été collectées, puis analysées. Nous avons ainsi découvert que, dans certains cas, où il y avait une variation formelle, le classement a été réalisé en fonction des aspects morphologiques ou compositionnels. De plus, il est possible que des sens figurés, symboliques ou métaphoriques aient été reconnus à partir d'une lecture plutôt culturelle. Nous avons donc trouvé des classements logiques s'éloignant des classements sémantiques attendus.

Par exemple, pour la référence **noyollo** (*mon cœur*), où une association avec des termes tels que **nomah** (*ma main*) ou **noyoliknih** (*mon ami chéri*) était attendue, il y a eu une forte préférence pour le terme **yoli** (*vivre*). Cela s'explique parce que ce dernier a été associé à son origine étymologique (**yol-**) et donc à un certain sens originel du mot **yollotl** (*cœur/ce qui a de la vivacité*).

Parfois les annotateurs ont fait appel à leurs compétences de bilinguisme en espagnol. Tel est le cas de **nemi** (*habiter/marcher*), qui a eu tendance à être associé à **yoli** (*vivre* en espagnol=*habiter*) plutôt qu'à **chantia** (*résider*). Parfois on a associé des mots pouvant composer une phrase pleinement significative : **tlahtolli** (*mot/récit/langue*) + **onikkak** (*je l'ai écouté*) pour composer **onikkak tlahtolli**, (*j'ai écouté un récit*), ou **noyollo + paki** (*être heureux*) pour construire **noyollo paki** (*mon cœur est heureux*), une salutation habituelle. Ainsi, on trouve des rangs (ou classements) qui montrent une différence culturelle entre ce qui a été compris comme : 1/ une association sémantique logique ; 2/ une extension des significations à partir d'usages culturels ; ou 3/ une interférence sémantique dans un contexte bilingue. Ce dernier type d'association n'était pertinent que pour certains termes bien spécifiques.

Les références elles-mêmes n'avaient pas de signification complexe, mais il fallait un profil de locuteur connaissant d'autres variétés et graphies pour réaliser la tâche de manière satisfaisante. Ainsi, le profil des annotateurs a été établi en tenant compte de cette condition. Il s'agit donc d'universitaires et d'étudiants master [14] qui utilisent le nahuatl à l'oral et à l'écrit dans le cadre de leurs activités professionnelles, de formation et de communication, et avec une connaissance relative, oral ou écrite, d'autres variantes géographiques et historiques.

3.2 Accord entre annotateurs et rang par consensus

Pour évaluer quantitativement l'accord des annotateurs sur le corpus pilote, on a utilisé le coefficient de concordance de Kendall, le κ de Fleiss [15] et l'entropie de Shannon. Le coefficient W de Kendall (Kendall & Smith, 1939) évalue 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. C'est le cas aussi pour le κ de Fleiss, qui est une extension du κ de Cohen (Fleiss & Cohen, 1973). L'entropie H (Shannon, 1948) a été utilisée pour déterminer l'incertitude ou l'hétérogénéité d'un ensemble. Dans un ensemble de rangs, elle renseigne sur leur cohérence ou leur diversité. Nous avons constaté que les deux mesures montrent une corrélation au sujet du – relativement – faible accord entre les annotateurs. Or, selon l'état de l'art, le W de Kendall est la métrique la plus adaptée pour estimer l'accord entre plusieurs rangs. Nous nous sommes donc

14. Maestriah ipan Totlahtol iwan Tonemilis/Maestría en Lengua y Cultura Nahua : https://www.uv.mx/mlcn/
15. https://fr.wikipedia.org/wiki/Kappa_de_Fleiss

concentré sur cette métrique : elle varie de 0,190 pour **melawak** (*correct*) à 0,598 pour **noyollo** (*mon cœur*). L'Annexe 1 montre un exemple des calculs de W, κ, H, et l'accord pour chaque référence.

En utilisant la méthode de Borda Count [16] sur le corpus pilote, nous avons construit un Rang par consensus (RC) représentant la majorité des rangs par référence (voir Annexe 2). Ce consensus a été utilisé principalement : 1/ pour estimer l'éloignement d'un annotateur i par rapport à l'ensemble des m annotateurs ; et 2/ pour évaluer la performance (via corrélation τ de Kendall) des différents ML. Il sera donc l'étalon par rapport auquel les différents modèles seront testés.

4 Résultats

4.1 Modèles de langue utilisés

Un ML est un outil conçu pour traiter et représenter les langues humaines. Au cœur de ces modèles réside l'utilisation des représentations vectorielles de mots, également appelées représentations denses, indispensables pour saisir les significations et les relations entre les mots dans un format adapté aux machines. Les représentations denses offrent un moyen puissant d'encoder à la fois des informations sémantiques et syntaxiques (Almeida & Xexéo, 2019). Elles sont essentielles pour des applications nécessitant une compréhension sémantique avancée, telle que la reconnaissance d'entités nommées, l'analyse de sentiments (Linhares-Pontes *et al.*, 2018) et la classification automatique des textes.

Dans notre étude, nous avons employé les ML statiques suivants : Word2Vec (Mikolov *et al.*, 2013), FastText (Bojanowski *et al.*, 2017) et Glove (Pennington *et al.*, 2014). Word2Vec et FastText avec leurs architectures CBOW et Skip-Gram, mettent en évidence les relations sémantiques basées sur les co-occurrences de mots, produisant ainsi des représentations vectorielles stables et intéressantes. FastText intègre des informations des sous-morphèmes (*subwords*), ce qui le rend particulièrement efficace pour le traitement des langues morphologiquement riches et agglutinantes. Il est performant aussi pour traiter les termes rares. Nous avons entraîné les ML statiques sur le corpus π-YALLI avec les hyperparamètres suivants : dimension D=(50, 100, 150, 200, 300), taille de la fenêtre $w = 5$ (valeur de la fenêtre ayant eu la meilleure performance) et époques=20 (la valeur où l'erreur d'apprentissage devenait stable). Étant donné un ML de dimension D, le calcul de similarité établit pour chaque référence i, un rang des j termes candidats au moyen de la distance cosinus entre l'embedding de la référence $\vec{R}_i$ et ceux des candidats $\vec{c}_j$, $sim(\vec{R}_i, \vec{c}_j) = \cos(\vec{R}_i, \vec{c}_j); i = 1, 2, ..., 23, j = 1, 2, ..., 5$.

D'un autre côté, les GML génératifs sont très puissants, mais aussi bien plus gourmands en ressources et également plus opaques que les modèles statiques. Nous avons testé huit GML en ligne bien connus : Mistral Large [17], DeepSeek-V3 [18] ; ChatGPT GPT4-mini [19], Gemini 2.0 [20], Copilot [21], Grok 3 [22], et Claude 3.7 Sonnet [23] ; ainsi que Llama-3.3-70B-Instruct [24] via HuggingFace. Le prompt structuré pour tous les GML a été le suivant : *« Étant donné le mot en nahuatl :* {"REF"}, *trier sémantiquement*

16. https://en.wikipedia.org/wiki/Borda_count
17. https://chat.mistral.ai/chat?q= Consulté le 19/02/25, avec environ 7,3 milliards de paramètres.
18. https://chat.deepseek.com/sign_in Consulté le 19/02/25, doté d'environ 200 milliards de paramètres.
19. https://chatgpt.com/ Consulté le 25/02/25 avec quelques dizaines de milliards de paramètres.
20. https://gemini.google.com Consulté le 25/02/25, nb de paramètres secret.
21. https://copilot.microsoft.com Consulté le 26/02/25, nb de paramètres secret.
22. https://x.ai/grok Consulté le 26/02/25, nb de paramètres secret.
23. https://claude.ai Consulté le 27/02/25, nb de paramètres secret.
24. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct 70 milliards de paramètres.

du plus proche au plus éloigné, les cinq mots suivants : {"CAND$_1$", "CAND$_2$", ..., "CAND$_5$"} ». Nous avons protégé "REF" et "CAND$_i$" par des guillemets afin d'éviter que les GML ne changent, n'introduisent, n'éliminent des caractères. Chaque GML fournit un rang qui sera utilisé par la suite.

Pour évaluer la performance des modèles dans la tâche sémantique proposée, nous avons calculé le coefficient de corrélation τ de Kendall (compris entre -1 et +1) entre RC et les rangs produits par les ML. Les modèles statiques ont été ré-entraînés localement sur le corpus π-YALLI, en utilisant soit la version avec l'unigraphie standardisée par règles, soit sans standardisation. L'hypothèse de base est que l'unigraphie standardisée pourrait permettre d'augmenter l'occurrence des *tokens*, de mieux les identifier contextuellement et d'augmenter ainsi la précision des modèles produits. Ceci a effectivement été confirmé lors de nos expériences.

Les résultats de la performance τ sont présentés sur la Table 3. À gauche les résultats des GML montrent une performance de 0,696 pour Gemini 2.0, qui a obtenu la plus haute performance. Il a aussi été le plus déterministe dans ses réponses en ligne, ce qui est un point non négligeable par rapport aux autres GML. À droite les modèles statiques entraînés sur le corpus π-YALLI sans unigraphie (en haut) et standardisé avec des règles d'unigraphie (en bas). FastText s'avère être le modèle le plus compétitif à 300 dimensions avec unigraphie, ayant une performance de **0,469**. La Table 4 montre la même tendance du comportement des modèles, estimée par $\bar{\tau}$ vis-à-vis des 27 annotateurs pris individuellement, ce qui permet de calculer la moyenne et l'écart-type.

Grands Modèles de Langue

GML pré-entraînés	τ
ChatGPT-4 mini	0,487
Claude 3.7	*0,644*
Copilot	0,399
DeepSeek V3	0,522
Gemini 2.0	**0,696**
Grok 3	0,452
Llama-3.3-70B-Instruct	0,107
Mistral Large	0,226

Embeddings locaux entraînés, époques=20, w=5

D	**Word2Vec**		**FastText**		**Glove**
	CBOW	Skip-Gram	CBOW	Skip-Gram	
	τ ; π-YALLI diverses graphies				
50	0,223	0,293	0,322	0,299	0,197
100	0,191	0,249	0,351	0,362	0,171
150	0,171	0,206	0,336	0,304	0,197
200	0,258	0,223	0,354	0,296	0,197
300	0,180	0,232	0,362	0,296	0,171
D	τ ; π-YALLI unigraphie standardisée				
50	0,220	0,212	0,368	0,328	0,273
100	0,246	0,278	0,399	0,394	0,325
150	0,244	0,252	*0,429*	0,336	0,336
200	0,287	0,339	*0,429*	0,362	0,359
300	0,269	0,278	**0,469**	0,365	0,339

TABLE 3 – τ de Kendall : GML et embeddings statiques vs. le Rang par consensus RC.

4.2 Analyse et discussion

La moyenne du coefficient de Kendall $\bar{W} = 0,389$, Kappa de Cohen $\kappa = 0,141$ et entropie $\bar{H} = 0,215$, montrent un certain degré de désaccord de l'ensemble d'annotateurs. Nous constatons que certaines références ayant des valeurs basses de Kendall, Cohen et d'entropie n'ont pas obtenu les classements attendus. En effet, quelques choix de classement montrent que certains termes candidats n'ont pas été clairement identifiés. Cela peut s'expliquer en raison de leur utilisation très locale qui a donc faussé les évaluations d'association sémantique. Il en va de même pour certains mots qui ont pu être considérés comme des archaïsmes ou qui étaient inconnus. Dans d'autres cas, le

Grands Modèles de Langue

GML	$\bar{\tau}$	GML	$\bar{\tau}$
ChatGPT-4 mini	0,277 ± 0,138	Gemini 2.0	**0,380** ± 0,170
Claude 3.7	0,371 ± 0,159	Grok 3	0,268 ± 0,121
Copilot	0,214 ± 0,108	Llama-3.3-70B-Instruct	0,070 ± 0,095
DeepSeek V3	0,289 ± 0,123	Mistral Large	0,109 ± 0,078

Embeddings locaux entraînés, époques=20, w**=5**, π-YALLI unigraphie standardisée

D	**Word2Vec** $\bar{\tau}$		**FastText** $\bar{\tau}$		**Glove** $\bar{\tau}$
	CBOW	Skip-Gram	CBOW	Skip-Gram	
50	0,140 ± 0,107	0,143 ± 0,087	0,179 ± 0,098	0,170 ± 0,112	0,166 ± 0,096
100	0,165 ± 0,105	0,186 ± 0,105	0,197 ± 0,107	0,203 ± 0,100	0,182 ± 0,083
150	0,164 ± 0,106	0,168 ± 0,090	0,207 ± 0,107	0,186 ± 0,091	0,176 ± 0,100
200	0,170 ± 0,117	0,196 ± 0,103	0,214 ± 0,104	0,187 ± 0,096	0,202 ± 0,101
300	0,177 ± 0,110	0,133 ± 0,094	**0,226** ± 0,125	0,203 ± 0,099	0,191 ± 0,097

TABLE 4 – $\bar{\tau}$ de Kendall : GML et embeddings statiques vs. les rangs de 27 juges.

classement semble avoir été fait sur la base d'une logique syntaxique, selon laquelle les verbes sont appariés avec des substantifs pouvant être objet ou sujet. Une basse valeur W témoigne peut-être de méconnaissances de graphies ou de mots peu utilisés parmi les communautés nahuaphones, et en particulier pour les annotateurs ayant participé à la tâche.

Par rapport à la similitude sémantique, les résultats sur la Table 3 montrent que, malgré le fait que deux GML performent mieux que les ML statiques ré-entraînés sur le corpus π-YALLI, ces derniers présentent un certain nombre d'atouts. En particulier, les ML statiques sont :

1. bien plus explicables : leur nombre de paramètres est faible et connu, et la similarité sémantique est calculée clairement ;
2. stables : les résultats ne changent pas en fonction du temps, ce qui arrive souvent aux GML en fonction du mode d'accès et/ou des paramétrages ;
3. pas d'hallucinations : phénomène très fréquent dans les GML qui peut avoir des répercussions considérables ;
4. plus économiques et donc plus écologiques que les GML.

Or, il était prévisible que les GML aient une meilleure performance que les ML statiques, car il s'agit d'une tâche de similarité sémantique mono-terme, où les termes sont très probablement présents dans des dictionnaires (nahuatl → espagnol/français/anglais). De ce fait, les GML cherchent plutôt de similitudes au sein des définitions présentes dans de langues autres que le nahuatl. Or, nous pensons qu'ils échoueraient dans des tâches de similarité entre phrases, car les phénomènes d'agglutination et de polysynthèse risquent de compromettre leurs performances. En revanche, les ML statiques cherchent des similarités contextuelles directement dans les embeddings de termes nahuatl. Ils restent donc des modèles bien plus écologiques, économiques, solides et explicables que les GML. Cela dit, FastText a battu Llama (qui n'a pas eu accès à Internet), Mistral, Grok et Copilot, tout en étant à un niveau comparable à celui de ChatGPT et DeepSeek [25]. Ceci montre la compétitivité des embeddings statiques et l'apport positif de l'unigraphie dans le traitement des π-langues, vis-à-vis des GML complexes, chers et trop gourmands en termes d'énergie et de ressources informatiques.

25. Les modèles en ligne ont eu, peut-être, accès à Internet pour résoudre cette tâche sémantique, mais nous ne pouvons pas en être sûrs. Nous pouvons faire l'hypothèse que, si cet accès leur était interdit, leurs performances auraient été inférieures.

Un exemple des hallucinations que nous avons rencontrées lors de la consulte des GML concerne le mot **acontle** (jarre). Le système ChatGPT GPT4-mini a répondu : « *...en nahuatl signifie "gorge". Maintenant, trions les mots donnés en fonction de leur proximité sémantique avec "acontle" :*

- *ahko → En haut, au-dessus (proche, car la gorge est une partie du haut du corps).*
- *atl → Eau (relativement proche, car l'eau passe souvent par la gorge lors de la consommation).*
- *tekomatl → Crâne, tête (moins proche, mais lié anatomiquement à la gorge).*
- *komitl → Jarre, récipient (éloigné, sauf si l'on fait une analogie entre la gorge et un contenant).*
- *tlaonilli → Chose étrange, mystérieuse (le plus éloigné, car aucun lien direct avec la gorge). »*

Donc, ChatGPT a halluciné et confondu : **acontle** (jarre) → *gorge* ; **tekomatl** (jarre) → *crâne* quand il est composé de cheveux (*tzontli*), donc *notzontekon* (ma tête) et **tlaonilli** (boisson) → *chose étrange* ; ce qui biaise complètement la réponse.

5 Conclusions et travaux futurs

Cette étude montre que les modèles statiques sont encore très compétitifs pour traiter les π-langues car ils ont battu plusieurs GML dans la tâche de similitude sémantique de mots. Leur rapidité, explicabilité, stabilité et fiabilité sont des atouts non négligeables dans des tâches TAL spécifiques.

Bien que le corpus π-YALLI ait une taille réduite vis-à-vis de corpus τ-langues, nous pensons qu'il s'agit d'une ressource intéressante pour modéliser le nahuatl (Torres-Moreno *et al.*, 2024a). On pourra par exemple, étudier l'impact de sa taille dans l'apprentissage – profond ou pas – des Modèles de Langue nahuatl. Nous augmentons par ailleurs constamment le volume du corpus π-YALLI. Ce corpus permettra de développer des outils d'analyse TAL classiques, d'améliorer nos ML statiques et probablement de générer des Mini-GML, des outils que nous disséminerons à la communauté scientifique. De plus, l'utilisation émergente et croissante du nahuatl (Figueroa-Saavedra, 2024) dans les réseaux sociaux, dans l'édition, dans les études universitaires et dans la diffusion scientifique — recherche d'information, reconnaissance d'entités nommées, résumé automatique (Torres-Moreno, 2014) -– rend ces outils de plus en plus nécessaires pour l'accès et la gestion de l'information numérique. Cette accessibilité permettra de relier différentes communautés nahuaphones situées dans des régions et des pays différents, ainsi que de faire circuler les connaissances exprimées dans cette langue auprès des étudiants et des spécialistes. Il s'agit donc d'un élan puissant pour faire mieux connaître cette importante π-langue. Il est prévu d'indexer une version enrichie d'annotations grammaticales et de nouvelles métadonnées, dès que les outils seront disponibles. Cette version devra permettre des requêtes selon les lemmes ou les catégories grammaticales, en plus des critères déjà mentionnés. Une fois constitué, sous forme de texte brut avec métadonnées identifiant chaque texte, le corpus π-YALLI 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).

Pour augmenter la taille du corpus, nous utiliserons les données textuelles, lexicales et les informations grammaticales disponibles dans le site web du Wiktionnaire en nahuatl [26]. Également nous allons procéder à la numérisation des documents en support papier dont nous disposons. Puisqu'aucun modèle OCR nahuatl n'est actuellement disponible à notre connaissance, nous implémenterons des stratégies de mixture d'algorithmes OCR afin de reconnaître les termes nahuatl. À plus long terme,

26. https://nah.wiktionary.org/wiki/Pewalistli

la transcription automatique de vidéos en provenance des réseaux sociaux ou des plateformes de diffusion (mono ou bilingues), est aussi envisagée dans le cadre d'un élargissement de nos projets de recherche sur le nahuatl. Finalement, nous nous consacrerons au développement du ML **BERTL**, basé sur l'architecture BERT et un tokeniseur crée spécifiquement pour le Nahuatl, qui permettra d'obtenir des embeddings mieux contextualisés et dynamiques (Rust *et al.*, 2021).

Annexe 1 : Exemple du protocole d'évaluation

W **de Kendall.** Étant donné n éléments à classer et m rangs indépendants (nombre de juges) on calcule : $W = 12\sum^{n}(R_i - \bar{R})^2/[m^2(n^3 - n)]$ où R_i est la somme des rangs assignés à l'élément i et $\bar{R}$ la moyenne de ces sommes. W=1 indique une concordance parfaite où tous les classements sont identiques, et W=0 une absence totale de concordance (les positions dans les classements sont complètement incohérentes). Nous employons directement la valeur W qui est normalisée en [0,1].

Entropie. Cette statistique a été calculée avec : $H(x) = -\sum^{n} p_i \log(p_i)$ où p_i est la probabilité (ou fréquence relative) que l'élément x soit dans la position i. Les valeurs vont de 0 (cohérence totale entre les rangs) à la valeur maximale possible $H_{\max} = \log_2(n)$ (divergence totale entre les rangs qui a lieu quand toutes les positions sont équiprobables, $p_i = \frac{1}{n}$). Nous avons normalisé la sortie entre $[0, \log_2(n)]$, où n=5 correspond au nombre de candidats par référence de la tâche sémantique.

κ **de Fleiss**. L'indice κ de Fleiss (extension du κ de Cohen), est utilisé pour évaluer l'accord entre plusieurs annotateurs qui classifient un ensemble d'éléments en catégories mutuellement exclusives. Cette métrique considère l'accord espéré par hasard. Le calcul est fait avec $\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$, où $\bar{P}$ est l'accord moyen observé, c.a.d $\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$, où $P_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$ est le rapport d'accord pour annotateur i. $\bar{P}_e = \sum_{j=1}^{k} p_j^2$, est l'accord aléatoire et $p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$ le nombre de fois où la catégorie j a été assignée à l'élément i. Ainsi κ=1 représente un accord parfait, κ=0 un accord aléatoire, κ<0 accord inférieur à celui aléatoire.

Voici le protocole d'évaluation appliqué à la référence **tototl** (oiseau). Soient n=3 candidats et m=3 annotateurs $\{J_1, J_2, J_3\}$ ayant annoté le rang : **tototl** ; candidats = {*tepostototl* (avion), *kwawtli* (aigle), *koyotl* (coyote)} ; d'où $\bar{R}$=6 pour la référence **tototl**. Avec n=m=3 on obtient W=0,778. Nous avons calculé $H(x)$ pour chaque référence. Le candidat *tepostototl* occupe une fois la position 1 ($p_1 = \frac{1}{3}$) et deux fois la position 2 ($p_2 = \frac{2}{3}$). On obtient $H(tepostototl) \approx 0,918$. $H(kwawtli) \approx 0,918$ et $H(koyotl)$=0. La moyenne d'entropie $\bar{H} \approx 0,612$. On obtient donc plutôt un bon accord des annotateurs pour **tototl**. En utilisant le rang : **tamalli** (tamal) [27] ; candidats = {*sakawilli* (zacahuil) [28], *tlaxkalli* (tortilla), *nixtamalli* (nixtamal) [29]}, $\bar{R}$=6 pour **tamalli**, et W=0. L'entropie $H(sakawilli) \approx H(tlaxkalli) \approx H(nixtamalli)$=1,585, donc $\bar{H} \approx 1,585$, ce qui correspond à la valeur maximale $\log(3)$. W et $\bar{H}$ montrent donc un total désaccord des annotateurs pour **tamalli**.

	tototl			**tamalli**		
Juge	*tepostototl*	*kwawtli*	*koyotl*	*sakawilli*	*tlaxkalli*	*nixtamalli*
J_1	1	2	3	1	2	3
J_2	2	1	3	2	3	1
J_3	2	1	3	3	1	2
R_i	5	4	9	6	6	6

Pour calculer l'indice κ, on utilise le nombre de candidats $N = 3$, le nombre de juges $n = 3$, et le

27. [...] nom générique donné à plusieurs plats [...] d'origine indigène https://fr.wikipedia.org/wiki/Tamal
28. Une sorte de tamal mexicain, de grande dimension : https://fr.wikipedia.org/wiki/Zacahuil
29. https://fr.wikipedia.org/wiki/Nixtamalisation

nombre de catégories $k = 3$. On constitue une table avec les valeurs n_{ij} qui comptent le nombre de juges qui ont assigné le candidat i à la catégorie j, puis on calcule $\bar{P}$, $\bar{P}_e$ et κ.

	tototl			**tamalli**		
	tepostototl	*kwawtli*	*koyotl*	*sakawilli*	*tlaxkalli*	*nixtamalli*
1	1	2	0	1	1	1
2	2	1	0	1	1	1
3	0	0	3	1	1	1
P_i	0,5	0,5	0,5	0,5	0,5	0,5
	$P = 0,5$	$P_e = 0,333$	$\kappa = 0,25$	$P = 0,5$	$P_e = 0,333$	$\kappa = 0,25$

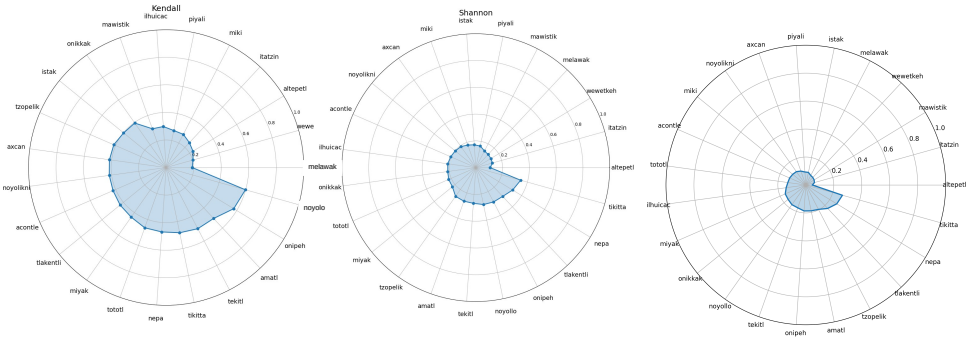

FIGURE 1 – W, H, κ, mesurant l'accord entre annotateurs selon les références.

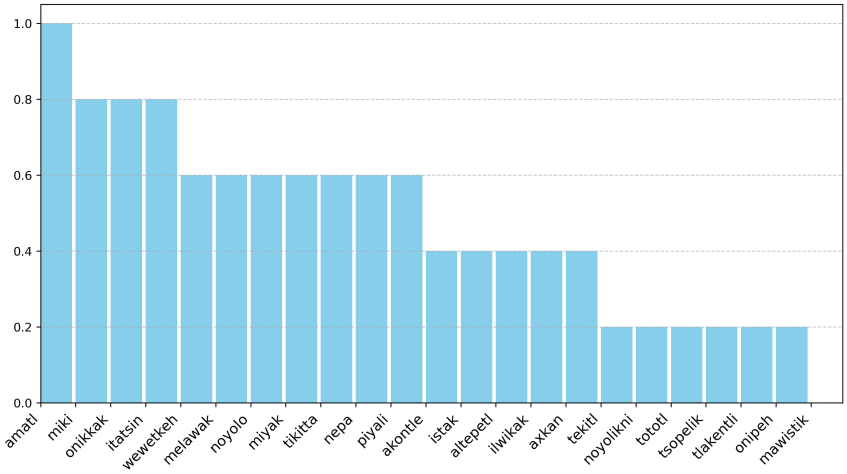

FIGURE 2 – τ de Kendall 1.8 π-yalli e=20, D=300, w=5, FastText-cbow.

Annexe 2 : Rang par consensus RC

i	REFERENCE	RANG PAR CONSENSUS RC				
		$candidat_1$	$candidat_2$	$candidat_3$	$candidat_4$	$candidat_5$
1	**wewetkeh (N)** les vieillards	**tokokoltzitzin (N)** nos grands-parents	**ilamantsij (N)** la vieille	**nouеueh (N)** mon mari	**weyi (N)** grand	**katka (V)** il était, il fut
2	**miki (V)** il meurt	**ixpoliwi (V)** il meurt, il disparaît	**mihkailwitl (N)** jour des morts	**kitoka (V)** on l'enterre	**kokostli (N)** maladie	**siwatl (N)** femme
3	**acontle (N)** jarre	**komitl (N)** assiette	**tekomatl (N)** jarre, vase	**atl (N)** eau	**tlaonilli (N)** boisson	**ahko (Loc)** au-dessus
4	**istak (N)** blanc	**chipawak (N)** blanc, pur	**astatik (N)** blanc brillant	**tlapalli (N)** couleur	**istatl (N)** sel	**itztli (N)** obsidienne
5	**tekitl (N)** travail	**tekipanolli (N)** travail, emploi	**titekitih (V)** nous travaillons	**nitlachihua (V)** je fais, je fabrique	**Tekillah (N)** Tequila (toponyme)	**tlakualli (N)** nourriture
6	**altepetl (N)** ville, village	**altepeyokan (N)** comarque	**chinanco (N)** village	**Coatepec (N)** Mont du serpent	**tepetl (N)** colline	**hueyatl (N)** mer
7	**noyolikni (N)** mon cher ami	**toikniwah (N)** nos amis	**towampoyowan (N)** nos collègues	**noyollo (N)** mon cœur	**niyolpaki (V)** je suis heureux	**naneh (N)** madame
8	**tototl (N)** oiseau	**kuawtli (N)** aigle	**patlani (V)** on vole	**tepostototl (N)** avion	**coyotl (N)** coyote	**miyak (N)** beaucoup
9	**ilhuicac (Loc)** au ciel	**sitlalpan (Loc)** au lieu des étoiles	**tlakpak (Loc)** haut, au-dessus	**Miktlan (N)** Lieu des morts	**tlalli (N)** terre, sol	**nikilwia (V)** je lui dis
10	**melawak (N)** correct	**kualli kah (Interj)** c'est bon	**yompa (Interj)** C'est correct	**ompa (Loc)** là-bas	**nimomati (V)** je me sens bien	**tekitl (N)** travail
11	**tzopelik (N)** sucré	**ahwiyak (N)** savoureux, aromatique	**pantzin (N)** brioche	**tamalli (N)** tamal	**chilatolli (N)** soupe de maïs piquante	**iztak (N)** blanc
12	**amatl (N)** papier, document	**amoxtli (N)** livre	**iswatl (N)** feuille de plante	**kuawitl (N)** bois, arbre	**aman (Loc)** maintenant	**axkan (Loc)** maintenant
13	**noyollo (N)** mon cœur	**yoli (V)** on vit	**noyolikni (N)** mon cher ami	**paki (V)** il est content	**momah (N)** ta main	**tototl (N)** oiseau
14	**mawistik (N)** étonnant	**tetzawitl (N)** prodige, monstre	**tlapowalli (N)** récit, conte, lecture	**tenyo (N)** célèbre	**mawissotl (N)** respect, admiration	**nimajmaui (V)** j'ai trop peur
15	**tlakentli (N)** linge	**kechkemitl (N)** quechquemil, sarape de femme	**wipilli (N)** huipil, chemisier	**cactli (N)** huarache, chaussure	**tlakayotl (N)** corps	**tlapowalli (N)** conte, récit
16	**miyak (N)** beaucoup	**miakeh (N)** plusieurs	**nochi (N)** tout	**seki (Quant)** quelques uns	**achi (Quant)** suffisant	**tlamantli (N)** chose
17	**onipeh (V)** j'ai commencé	**tipewa (V)** tu commences	**xikpewalti (V)** faites-le	**tlapewaloyan (N)** début	**achtopa (Loc)** introduction	**ontlami (V)** se termine
18	**tikitta (V)** tu le vois	**tikchiya (V)** tu le regardes	**tix (N)** nos yeux	**otikittakeh (V)** nous l'avons vu	**tlanesi (V)** il y a de la lumière, il fait	**tlakualli (N)** nourriture
19	**onikkak (V)** j'ai entendu	**xikkakikan (V)** écoutez-le	**nikahsikamati (V)** je comprends	**nonakas (N)** mon oreille/ouïe	**tlahtolli (N)** ce qu'on dit, parole	**yalwa (Loc)** hier
20	**itatzin (N)** son père	**notatzin (N)** mon père	**inantzin (N)** sa mère	**pilhuah (N)** celui qui a des fils	**Totiotzin (N)** notre Dieu	**motta (V)** ça se voit
21	**nepa (Loc,Dem)** là-bas, ça	**ompa (Loc)** là-bas	**inin (Dem)** cela	**oncan (Loc)** là	**ni (Dem)** cela	**komitl (N)** assiette
22	**axcan (Loc)** maintenant	**axan (Loc)** aujourd'hui, maintenant	**nama (Loc)** aujourd'hui, maintenant	**niman (Loc)** après	**tonalli (N)** ce jour	**yalwa (Loc)** hier
23	**piyalli (Interj)** salut	**panolti (Interj)** salut, allez-y	**tlanextili (Interj)** bonjour	**tateh (N)** monsieur	**asta mostla (Interj)** à demain	**xikpia (V)** tenez

TABLE 5 – Corpus du Rang par consensus RC de couples références-candidats (27 annotateurs, voir Section 3). Entre parenthèses les catégories grammaticales des mots : Interj=Interjection, Loc=Particule de localisation spatio-temporelle, Dem=démonstratif, N=substantif (certains noms, morphologiquement substantifs en nahuatl, sont traduits en français par les catégories adjectifs et pronoms), Quant=Particule quantitative, V=verbe.

Remerciements

Ces travaux ont été soutenus par la Structure Fédérative de Recherche Agorantic et par l'École Universitaire de Recherche InterMEDIUS (toutes les deux de l'Université d'Avignon), pour le financement du projet NAHU² et la thèse de J.-José Guzmán-Landa (Laboratoire Informatique d'Avignon, France) respectivement. Merci à tous nos relecteurs anonymes, dont leurs conseils nous ont permis d'améliorer cet article.

Références

ABDILLAHI N., NOCERA P. & 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.

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. (2019). Word embeddings : A survey. *arXiv :1901.09069*.

BERMENT V. (2004). *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"*. Thèse de doctorat, Université Joseph-Fourier - Grenoble I.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the ACL*, **5**, 135–146.

DE DURAND-FOREST J., DEHOUVE D. & ROULET E. (1995). *Parlons Nahuatl. La langue des Aztèques*. L'Harmattan.

EVERT S. & HARDIE A. (2011). Twenty-first century corpus workbench : Updating a query architecture for the new millennium. In *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.

FIGUEROA-SAAVEDRA M. (2023). Marcadores y conectores discursivos en la textualidad náhuatl entre universitarios nahuahablantes. *Cultura, Lenguaje y Representación*, **31**, 237–263.

FIGUEROA-SAAVEDRA M. (2024). *Amapowalistli iwan tlahkuilolewalistli. Tlamachtilamoxtli*. Universidad Veracruzana.

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

FLEISS J. L. & COHEN J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, **33**(3), 613–619. DOI : 10.1177/001316447303300309.

GUTIERREZ-VASQUES X., SIERRA G. & POMPA I. H. (2016). Axolotl : a web accessible parallel corpus for Spanish-Nahuatl. In *10th LREC'16*, p. 4210–4214.

INEGI (2020). : https://www.inegi.org.mx/rnm/index.php/catalog/632/study-description.

KENDALL M. G. & SMITH B. B. (1939). The problem of m rankings. (3), 275–287. DOI : 10.1214/aoms/1177732186.

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., *TALN. Volume 1 - Articles longs, articles courts de TALN*, p. 311–320, Rennes, France : ATALA.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA : MIT Press.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS - Vol 2*, NIPS, p. 3111–3119, Red Hook, NY, USA : Curran Associates Inc.

MOSELEY, CHRISTOPHER (ED.) & NICOLAS ALEXANDRE (CART.) (2012). *Atlas des langues en danger dans le monde*. UNESCO.

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.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *2014 EMNLP*, p. 1532–1543 : ACL.

RUST P., PFEIFFER J., VULIĆ I., RUDER S. & GUREVYCH I. (2021). How good is your tokenizer ? on the monolingual performance of multilingual language models. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 3118–3135, Online : ACL. DOI : 10.18653/v1/2021.acl-long.243.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

SMITH M. (2002). *The Aztecs*. Peoples of America. Wiley.

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

TORRES-MORENO J.-M., AVENDAÑO-GARRIDO M.-L., FIGUEROA-SAAVEDRA M., RANGER G., GONZÁLEZ-GALLARDO C., PONTES E. L., MORALES P. V., TORRES L. Q. & GUZMÁN-LANDA J.-J. (2024a). NAHU² : Un nouveau corpus pour le Nahuatl. In *18èmes Journées Informatique Région Centre-Val de Loire* : https://hal.science/hal-04814636.

TORRES-MORENO J.-M., GUZMÁN-LANDA J.-J., RANGER G., GARRIDO M. L. A., FIGUEROA-SAAVEDRA M., QUINTANA-TORRES L., GONZÁLEZ-GALLARDO C.-E., PONTES E. L., MORALES P. V. & JIMÉNEZ L.-G. M. (2024b). π-yalli : un nouveau corpus pour le nahuatl. *arXiv preprint arXiv :2412.15821*.