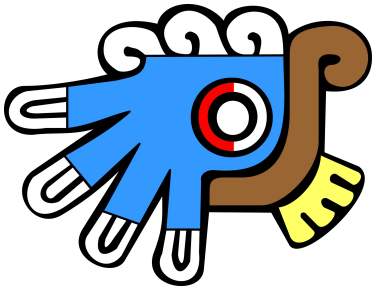
NAHU²

YANKUIK CORPUS PAMPA NAWATLAHTOLLI

{Juan-Manuel.TORRES, Graham.RANGER}@univ-avignon.fr

01.12.2025



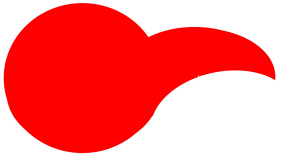
agorantic
Culture, Patrimoines et Sociétés Numériques
Fédération de Recherche

AVIGNON UNIVERSITÉ

cnrs

YANKUIK CORPUS PAMPA NAWATLAHTOLLI

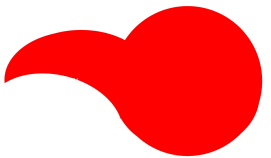
¡Piyalli!



Neh notoka Juan-Manuel

Nichantia Avignon Francia iwan niwalewa Veracruz, Mexihko

UN NOUVEAU CORPUS POUR LE NAWATL



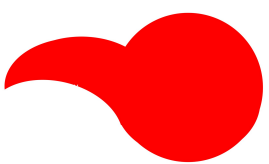
Bonjour !

Je m’appel Juan-Manuel

J’habite Avignon France et je viens de Veracruz, au Mexique

PART I

On utilise des mots nawatl, même sans le savoir…



Nawatl

Totopochtli ▶ **TOTOPOS**

Awakatl / Avagatl ▶ **AVOCAT**

Awakamolli ▶ **GUACAMOLE**

Koyotl ▶ **COYOTE**

Kakawatl ▶ **CACAHUATE**
CACAO

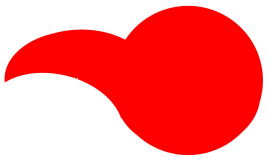
Xokolatl ▶ **CHOCOLAT**

Tomatl ▶ **TOMATE**

Chayotl ▶ **CHAYOTE**

Immersion nawatl[0,1]

Petit exercice linguistique



Quel est le mot le plus long en nawatl?...

[1] *Inspiré de Superholly : https://www.youtube.com/watch?v=C2taeD-qXlw*
[0] *Modification : Miguel Figueroa-Saavedra Universidad Veracruzana*

Nawatl - grammaire[2]

Prefixes + **Racine VERBE** + ***Suffixes***

- ***Prefixes***: SUJET | OBJET | ADVERBES…
- ***Suffixes***: TEMPS | MODE | …

(*VERBE + VERBE*)

[2] *Parlons Nahuatl. Durand-Forest et al, 1995, L'Hartmann, Paris*

Verbes

Qui a fait quoi? à qui? quand? l'a-t-il fait? vraiment?…

- Verbes « plongés » dans d'autres verbes
- Substantifs « verbalisés »

IX K A = *griller*

I X (i*sh*) **K** **A**

Mais on ne peut pas dire uniquement ***ixka***…

- **ixka** = «*griller*» (...mais quoi?) **Racine transitive**
- **Ni-k**-ixka = « **Je *le*** *grille* »
  - **Ni : Je** **k / ki : objet direct**
  - **tla :** objet non spécifié
  - Ni-**tla**-xka = « Je *grille* ***quelque-chose*** »
  - Tla-xka-**l** = « **Quelque-chose grillée** »

Tla-xka-l

«Quelque-chose grillée »

Tlaxkal-**li** = « Tortilla » (t)**li** : sustantif

Tla-xkal-li

Tlaxcala

« *Quelque-chose-grillée* »

Tlaxkal-**li** = « *Tortilla* »

- **tlan** = endroit
- Tlaxkal-(t)**lan** ► Tlaxkal-**lan** = *Tortilla* + *endroit*

► ***Tlaxcala***

Sustantif ▸ Verbe

- Suffixe **-owa** / **-oa** : « verbaliser » un sustantif
- Tlaxkal-**owa** = « *Faire-tortillas* » (processus)
- **Tlaxkalowa** = « Ce *qu'on fait à ce qu'on grille* »

Transformation en verbe !

- Mais on ne peut pas dire **tlaxkalowa**... il faut ajouter :

qui fait quoi? à qui? ...

Ni-k + Tlaxkalowa

Processus pour faire tortillas: donner des petites tapes

Ni-**k**-tlaxkalowa = «***Je lui*** *donne des petites tapes (applatir)*»

- **Ni** = Je
- **k** = 3eme personne singulier

- **Niktlaxkalowa**

Freepik.com

Ma+Tlaxkalolli

- Tlaxkal+**o**(wa)+**l** = **Tlaxkalo+l+**(t)**li** = « Ceci *à ce qu'on donne des petites tapes* »
- **Ma**+tlaxkalol-li
  - **MA(ITL)** = main / bras

- **Matlaxkalolli** = « Ceci *à ce qu'on donne des petites tapes, avec la* ***main*** »

Tlatlikuin

- **Tlatlikuin** = « *Son d'un coup qui resonne* » (V)
- **Tlatlikuini** = « *Tonnerre* »

Tlatlikuinal-li = « L'*objet qu'on fait sonner* » (N)

Matlakxcaloltlatlikuinalli

- **Ma-Tlaxkalol-li** + **Tlatlikuinal-(t)li** = **Matlaxkaloltlatlikuinalli :**
- « L’*objet qu’on fait sonner en lui donnat des petites tapes, avec la main* »

Tzotzona (verbe)

- **Tzotzona** = « *Jouer un instrument musical de cordes ou de percussion* »

Tlatzotzontli (substantif)

- **Tlatzotzon*tli*** = « Quelque instrument musical de percussion/cordes qu'on joue »

Matlaxkaloltlatlikuinal+ Tlatzotzon(tli)

- **In** + (ma-tlakxalol-tlatlikuinal-tlatzotzon) + **wan**
- **In** : *ils/elles* **(...)** **wan** : *substantif pluriel possedé*
- « *ils/elles* **(...)** *possèdent-plusieurs de ces-choses* »

Quelles choses

?

Inmatlaxkaloltlatikuinal tlatzotzonwan

Quelle choses?

- *Les instruments musicaux de percussion joués avec des petites tapes avec la main, ce qui fait qu'ils resonnent… les* ***tambourines****!*

Inmatlaxkaloltlatikuinaltlatzotzonwan

« ***ils/elles sont en train de jouer leurs tambourines*** »

Inma tlaxkalol tlatikuinal tlatzotzonwan

- Verbe: « *ils/elles jouent leurs tambourines* » + **owa**

- Prêt du français: « ***tambourine*** » + h = « *tambourine**s*** **»**

  in + (tambourineh) + wan = intambourinehwan

  « *ils/elles ~~sont en train de jouer~~ ont leurs tambourines* »

  ~ « *ils/elles **ont** leurs tambourines* » (les jouent-ils?)

Le mot le plus long en nawatl…?

◄ Inmatlakxaloltlatlikuinaltlatzotzonwan ►

- Allonger à l’aide de préfixes et suffixes… (37 lettres)
- Polysyntétique

Quelle est ~~le mot~~ la phrase la plus longue en nawatl? ~

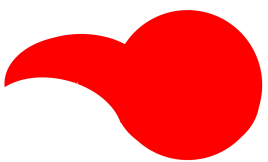
Quel est le mot le plus long en français (castillan)?

Anticonstitutionnellement (25 lettres) ***Electroencefalografista*** (23 lettres)

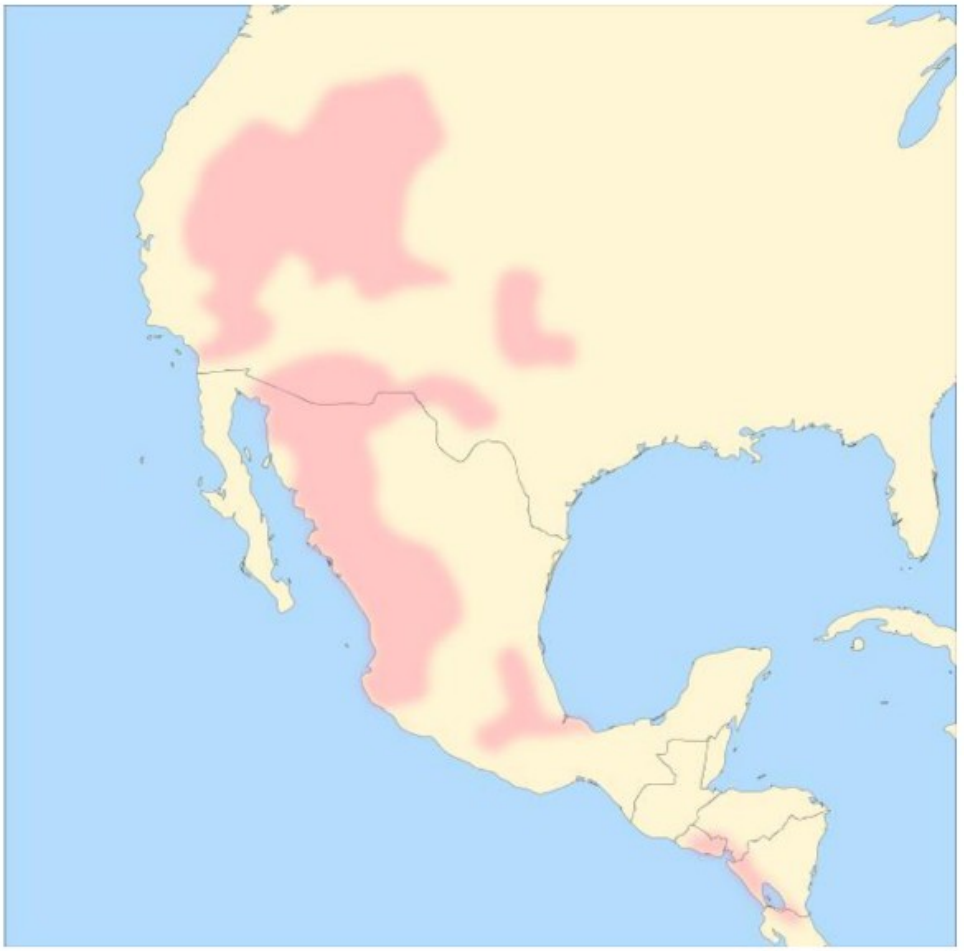
PART II

Commencer par le commencement…

les mots et les « mots-phrases »



TLAHTOLLI



[3] *Cours de Miguel Figueroa-Saavedra*
Université Veracruzana (Mexique)

- Protonahua
- Paleonahua- 1200 a.C.
- Neonahua- 800 d.C.

Nawatl[3]

Nawatl

2,5 millions de nahuaphones

Canada, EU, Maxique, Amérique Centrale

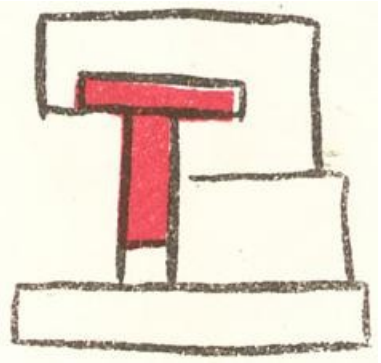
Langue nationale du Méxique

LANGUE VIVANTE - VARIETES EN DANGER

Nawatl - écrit

- Ecriture idéographique

- Ecriture alphabétique : plusieurs **graphies**
  - Alphabet franciscain
  - Alphabet jesuite
  - Alphabet traditionnel
  - Alphabet SIL
  - Alphabet pratique



Kalli = *Maison*

Tlahtoa, «parler»

- Tlahtoa
- Tlahtohua
- Tlatohua
- Tlatoa
- Tlahtowa
- Tlatowa
- Tlajtoa

Tlahtoa

Tlahtoani : « Celui qui parle » ~ (Roi, Gouverneur)

Kuauhtemoc

Unification de graphies

Unigraphie (perl, symbolique)

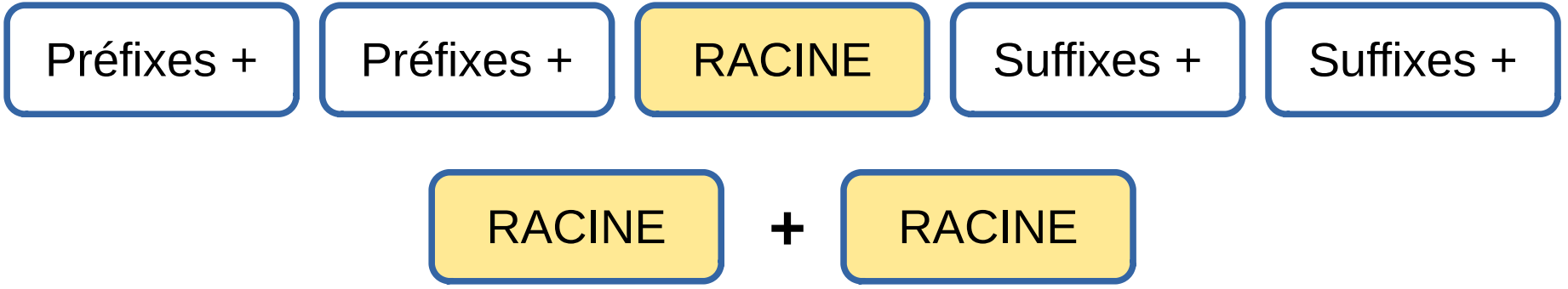
Rule	Pattern	Replacing
1	hu+vowel	hu→w
2	vowel+uh	uh→w
3	qua	q→k
4	que|qui	qu→k
5	(vowel|<sp>)+u+vowel	u→w
6	ca|co|cu	c→k
7	kw	w→u
8	(uk|ku)+vowel	c→k
9	c ∉ ch	c→k
10	consonant+y+letter	y→i
11	vowel+j+(vowel|consonant|<sp>)	j→h
12	ywan	ywan→iwan
13	<sp>+wan+<sp>	wan→iwan

Rule	Pattern	Replacing
14	yn	yn→in
15	ce | ci | çi | zi	[c ç z]→s
16	zo	z→s
17	tz | tc | tç	[c ç z] → s
18	ll | hl	(ll | hl) → l
19	<sp>+i+vowel	i→y
20	(<sp>|consonant)+u+consonant	u→o
25	(aa|ee|ii|oo)	→ a | e | i | o
26	(ha|he|hi|ho)	→ ah | eh | ih | oh
27	<sp>+(ahmo—hamo—ajmo)+<sp>	→ amo
28	vowel+c+hu+vowel	chu → kw
29-31-32	(ã|à|á|ā)|(è|é|ē)|(í|ī|ì)|(ò|ó|ō)	→ a | e | i | o
30	â‖ê‖î‖ô	→ ah ‖*eh*‖*ih*‖*oh*
33	(Ç ç) ‖($\tilde{y}\hat{y}\ddot{y}$)	→ s ‖*y*
34	ihua[n]	ihua[n]→ihuan paleography
35	vowel:word	a:tl→atl paleographie

Guzman et al. A symbolic algorithm for the unification of Nawatl word spellings, MICAI'25, pp 141-154

Nawatl

- Agglutinante - Polysynthéntique
- 15 consonnes (**ch h k kw l m n p t s ts tl x y *w***)
- 4 voyelles (**a e i o**)



Grammaire : composition

Composition nominale

- Toch(tli) + Kalli ► **Tochkalli** = *Tanière*

  Lapin + *Maison*

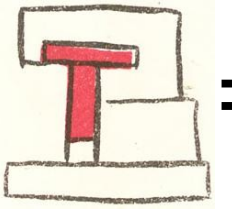
- Yollo(tl) + Tetl ► **Yollotetl** = vaillant

  Coeur + *Pierre*

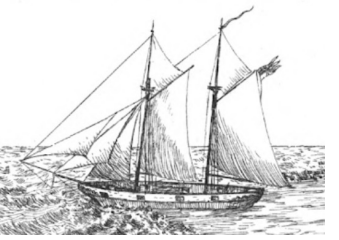
- A(tl) + Kalli ► **Akalli** = *bateau*

  Eau + *Maison/structure*

+



=



Langues π

Nawatl : dotée de ***peu*** *de ressources informatisées*: **π-langues**

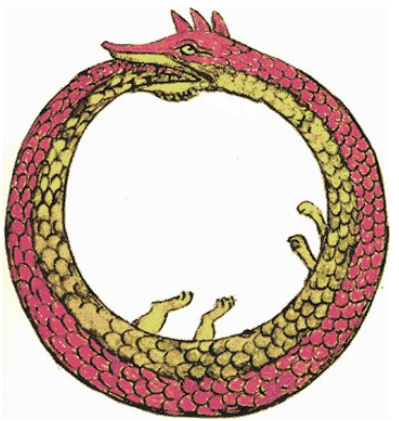
μ-langues (***moyennement*** dotées) **τ-langues** (***t****rès* bien dotées)

Pas d'analyseurs morphologiques

Pas de lemmatiseurs/stemmiseurs

Pas de modèles de langue (sérieux)

Traducteur Google 2024 (beta)



Pas de corpus suffisants…

π-langues

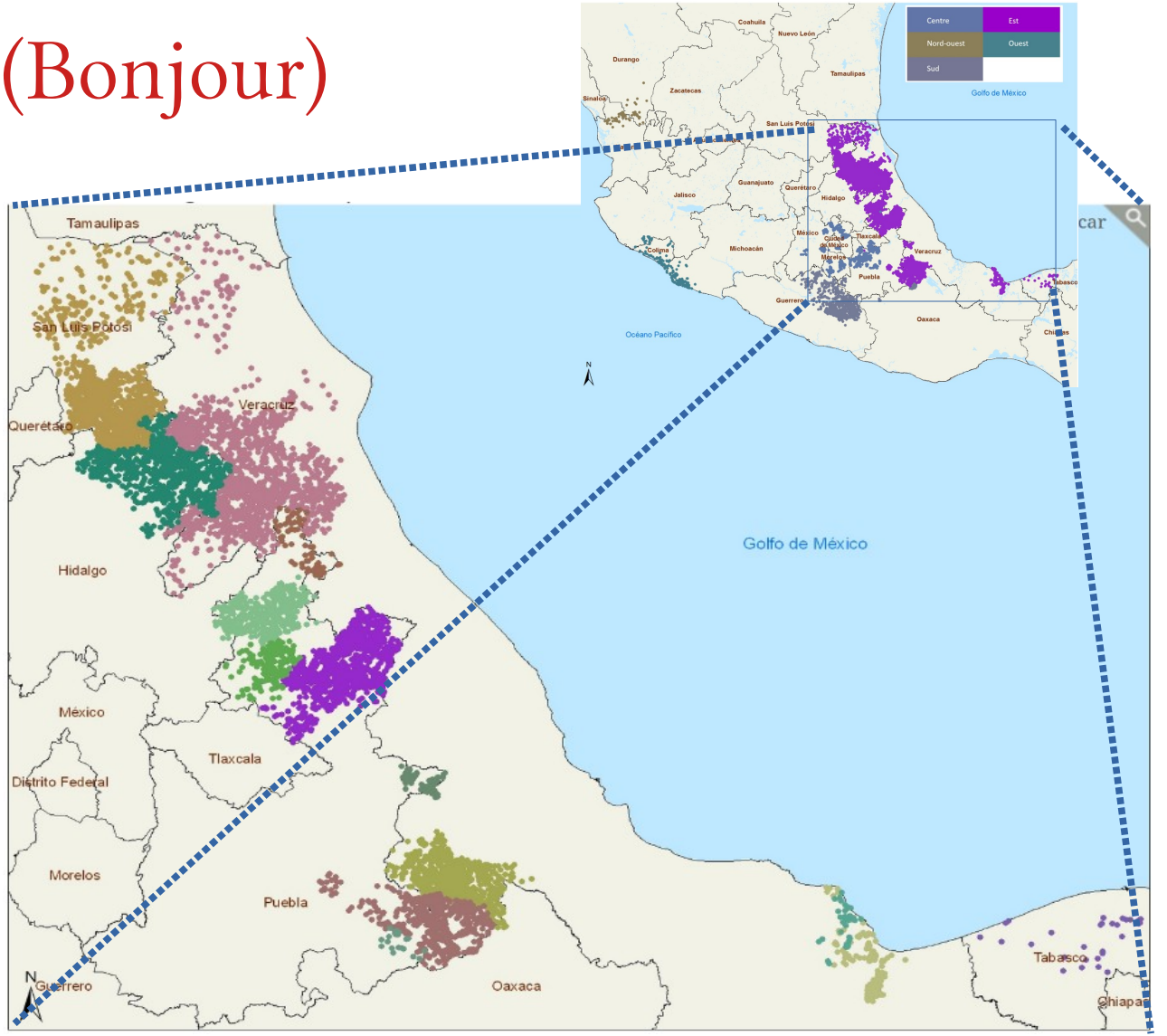
- **Ingrédients**
- **Outils**
- **Protocoles d'évaluation**

Thèse interdisciplinaire démarrée **10.2024**
(Informatique + Humanités numériques)

agorantic
Culture, Patrimoines et Sociétés Numériques
Fédération de Recherche
AVIGNON UNIVERSITÉ
cnrs

Corpus π-YALLI (Bonjour)

- **Hétérogène**
  - Documents historiques **HIS**
  - Littérature **LIT**
  - Thèses, Master **EDU**
  - Politique **POL**
  - Wiki, Blogs **WIK**
  - Science/Technologie **TEC**
  - **...**
- **Variétés**
  - Nawatl Central
  - Sur
  - La Huasteca
  - Classique, ...



Corpus π-YALLI[4]

- utf8
- + **53,3** Mo texte
- + **6,6** M poly-termes nawatl (multi-mots agglutinés, polysynthétiques)

Topic	Docs	Tokens	Sentences
AGR	3	7 828	251
COS	1	53 408	2 992
ECO	1	16 777	1 369
EDU	98	502 392	32 343
HIS	56	705 790	27 651
LEG	26	352 563	14 237
LIN	13	402 364	43 319
LIT	138	1 018 669	60 112
MED	4	14 250	736
MUS	5	4 306	408
PHR	49	9 259	1 238
POE	12	6 604	398
POL	3	1 800	68
REL	31	3 311 474	232 848
TEC	3	27 838	164
WIK	4 298	194 292	9 498
TOTAL	**4 746**	**≈ 6 629 000**	**≈ 428 000**

[4] *Guzman et al. 2025, pi-yalli: Un nouveau corpus pour le Nahuatl, TALN 2025, pp 802-816*

Corpus π-YALLI

- Adéquat : **IA classique** ✅
- Adéquat : **apprentissage d'*embeddings*** ✅
- ***Insuffisant*** *pour transformateurs IA* ***générative*** ❌

Qué faire? Commment l'augmenter ?

- OCR + Wiki + Blogs + textos → ***lent et pénible***...
- Grámmaires non contextuelles CFG → ***signification*** ?

Context-Free Grammar

Analyse

Génération...

- V is a finite set of **non-terminal symbols**.
- Σ is a finite set of **terminal symbols**, such that $V \cap \Sigma = \emptyset$.
- R is a finite set of **production rules**, of the form: $A \to \alpha$ where $A \in V$, $\alpha \in (V \cup \Sigma)$, and $(V \cup \Sigma)$ represent all possible strings of length ≥ 0, formed with symbols from V and Σ.
- $S \in V$ is the start symbol.

μgnaw$\oplus$0

Micro CFG *naïve*, basée sur langues indo-européennes (ES, FR)

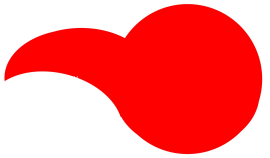
Non recursive

$P \rightarrow ADV_T \ (N|V)$
$N \rightarrow ADJ \ (ART\_|POS)\oplus n$
$V \rightarrow N \ NEG \ PV_3\oplus v \ ADV_Q$
$V \rightarrow PP_i \ NEG \ PV_j\oplus v \ ADV_Q \ ; i,j=1,2,3; i=j$

$ADV_Q \rightarrow$ miyak | tlawel | $\emptyset$ # a lot | too much | $\emptyset$
$ADV_T \rightarrow$ naman | axcan | axan | $\emptyset$ # now | this day | today | $\emptyset$
ADJ $\rightarrow$ tomawak | kualtzin | $\emptyset$ # fat | nice | $\emptyset$
ART $\rightarrow$ se | ni | $\emptyset$ # one | the, this | $\emptyset$
POS $\rightarrow$ no | mo | i # my | your | his, her, its
$PP_i \rightarrow$ na | ta | ya # I, me | you | he, she, it
$PV_j \rightarrow$ ni | ti | $\emptyset$ # I | you | he, she, it | $\emptyset$
NEG $\rightarrow$ amo | axkeman | $\emptyset$ # no | never | $\emptyset$

n $\rightarrow$ siwatl | miston | elotl | xokotl | tochin |
yolkatl | nakatl | $\cdots$
woman | cat | corn | fruit | rabbit | animal | meat | $\cdots$
v $\rightarrow$ nehnemi | kwa | kaki | $\cdots$
to walk | to eat | to listen | $\cdots$
$\oplus$=concatenation $\emptyset$=null _=space

Nawatl: structure basée sur le verbe



Structure	Exemple	Traduction
VSO	Kitta tlakatl kalli	voir (un) homme (une) maison
VO	Kitta kalli	(il/elle) voit (une) maison
VS	Kitta tlakatl	voir (leur) (un) homme
VOS	Kitta kalli tlakatl	voir (une) maison (un) homme
SV	Tlakatl kitta	(un) homme voir (leur)
SVO	Tlakatl kitta kalli	(un) homme voir (une) maison
SOV	Tlakatl kalli kitta	(un) homme (une) maison voir

μgnaw$\oplus$1

Micro CFG réaliste basée sur les structures nawatl

Non recursive

MT *marqueur temporel*
MIR *marqueur d'intensité relatif*
MIA *marqueur d'intensité absolu*
MV *intensité verbale*
ML *intensité de lieu*
MO *intensité d'objet*

→

CORPUS AUGMENTE !

$P \rightarrow VSO \mid \overrightarrow{VO} \mid \overleftarrow{VS} \mid VOS \mid SV \mid SOV \mid SVO$
$V \rightarrow NEG\ MT\ MIV\ MV_i \oplus MO_j \oplus v\ ;\ i\ ,\ j = 1\ ,2\ ,3\ ;\ i \neq j$
$S \rightarrow MCS\ ADJ\ POS \oplus n$
$O \rightarrow ADJ\ POS \oplus n\ ML$

$ADJ \rightarrow$	weyi | istak | $\emptyset$	# big | white | $\emptyset$
$POS \rightarrow$	no | mo | i | $\emptyset$	# my | your | his | $\emptyset$
$NEG \rightarrow$	amo | $\emptyset$	# no | $\emptyset$
$MT \rightarrow$	aman | cemicac | $\emptyset$	# now | forever | $\emptyset$
$MIV \rightarrow$	miyak | nochi | $\emptyset$	# a lot | all | $\emptyset$
$MCS \rightarrow$	san | miakpa | $\emptyset$	# only | often | $\emptyset$
$ML \rightarrow$	nikan | nepa | $\emptyset$	# up | there | $\emptyset$
$MV_i \rightarrow$	$\emptyset$	# he , she
$MO_j \rightarrow$	ki	# from him , her

v → toka | itta | chihua | pia | maka | neki | ···
bury | see | do | have | give | want | ...
n → siwatl | miston | elotl | xokotl | tochin |
yolkatl | nakatl | ···
woman | cat | corn | fruit | rabbit | animal | meat | ···
$\oplus$ = concatenation $\emptyset$ = null

Protocole d'evaluation

- **Tâches classiques TALN**
  - Similarité semántique (mots / phrases)
  - *Evaluation qualité : Kendall + évaluation humaine*
- **Algorithmes**
  - Word2vec ; FastText ; GloVE
  - LLM commerciaux

Les grands modèles LLM…

ChatGPT

Gemini

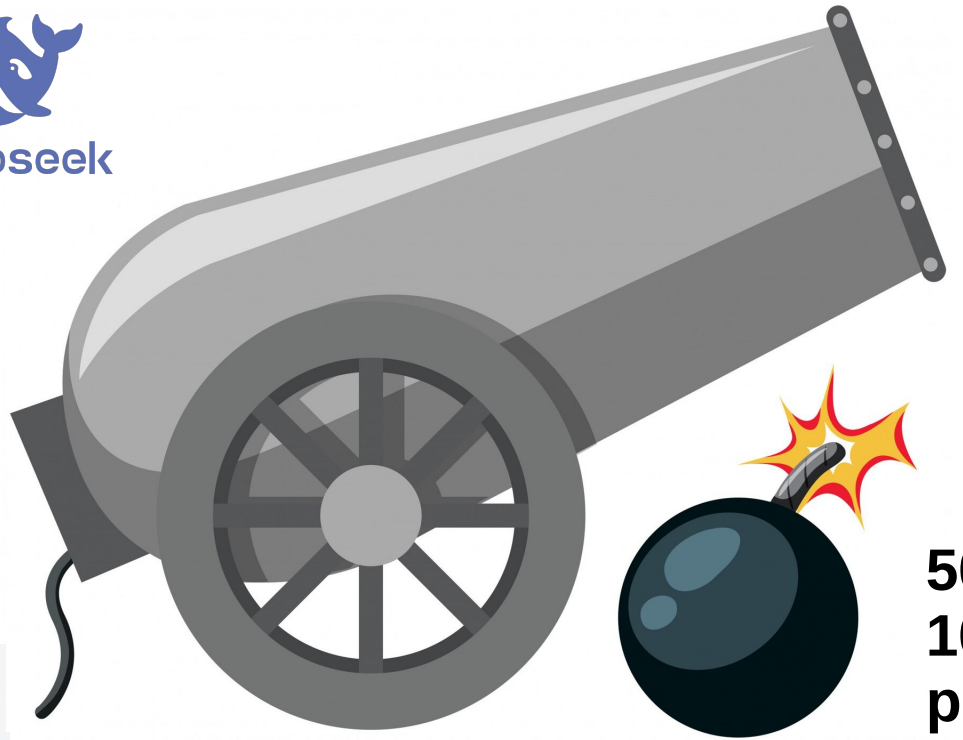
deepseek

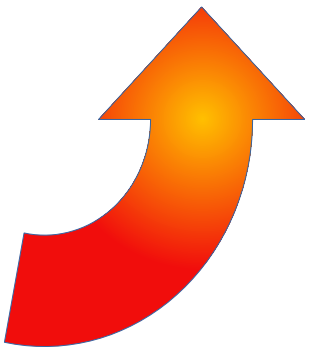
Grok

Copilot

MISTRAL AI_

Meta Llama 3





500 milliards...
1000 milliards de paramètres!

LLM: combien coûte leur entraînement?

deepseek

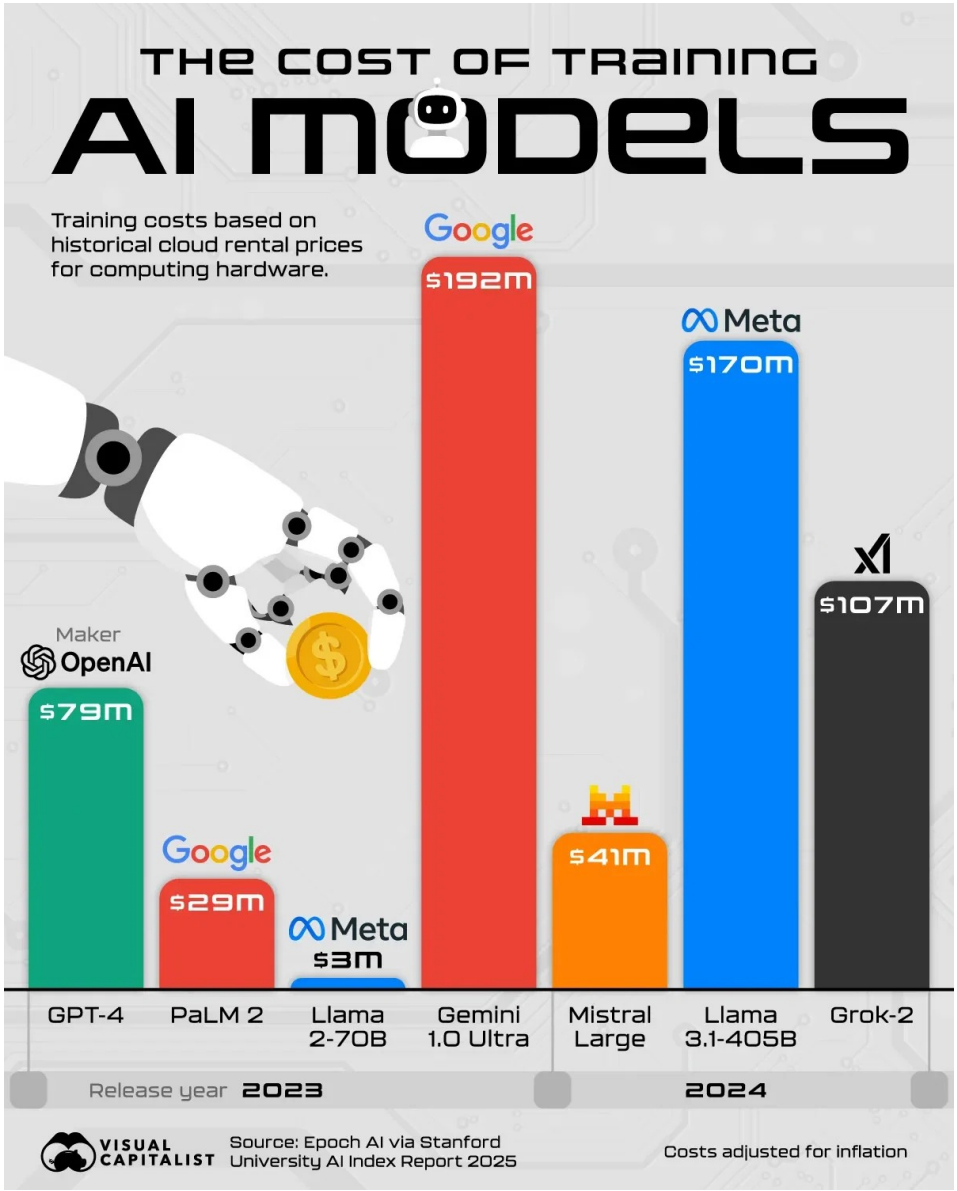
Reuters World Business Markets Sustainability Legal Commentary

China's DeepSeek says its hit AI model cost just $294,000 to train

By Eduardo Baptista

September 19, 2025 4:09 AM GMT+2 · Updated September 19, 2025

https://www.voronoiapp.com



Et les (*vieux et pas chers*) *embeddings* statiques...?

- **Word2Vec**
- **FastText**
- **Glove**

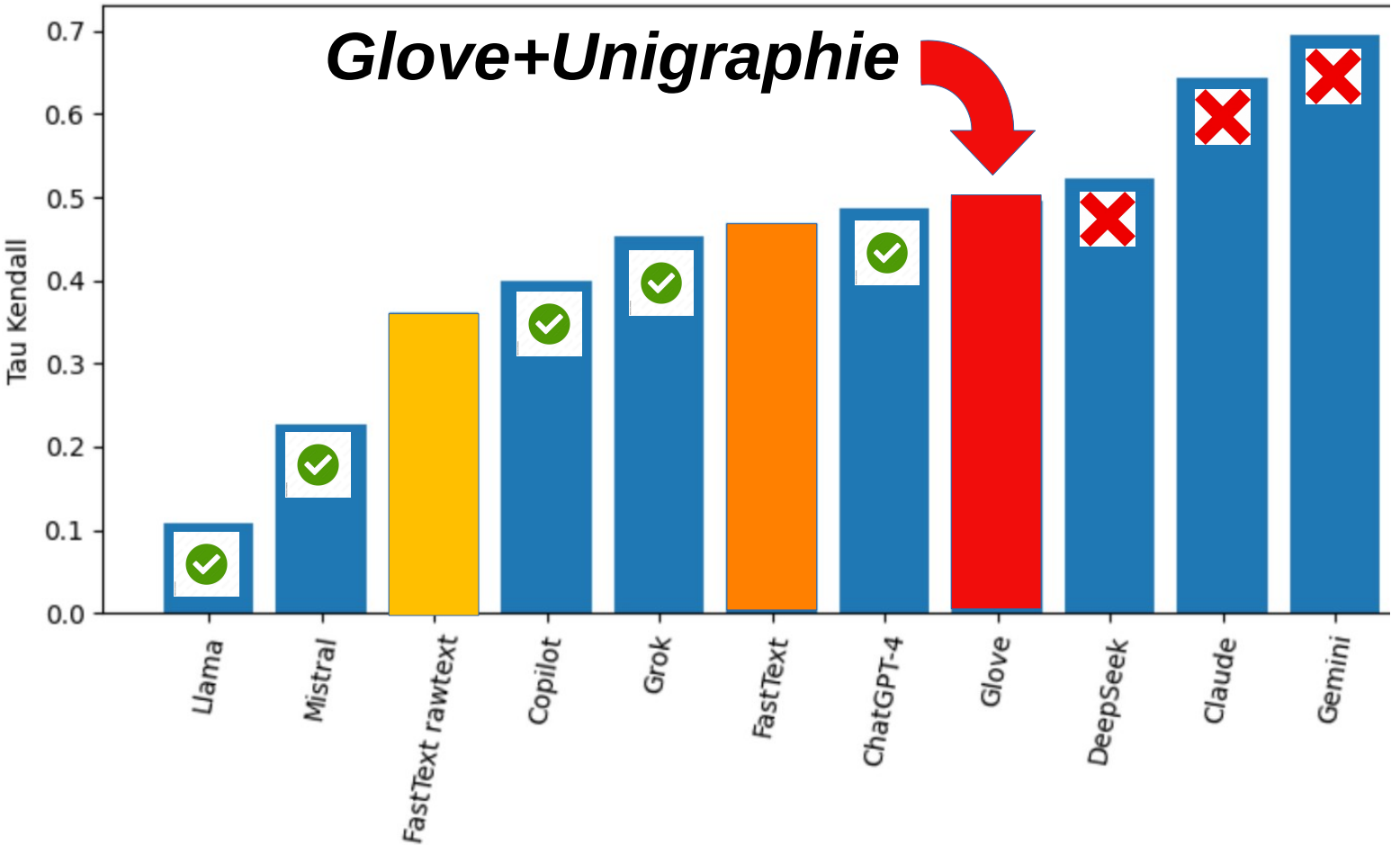
Tâche sémantique (mots)

- **23** térmes de **référence**
- **5 candidats** ayant sémantique variable vis-à-vis la **référence**
- **27 annotateurs** : **rang** de 5 candidats pour les références

- **Exemple** (sémantique proche ► lointaine)

TOTOTL :	**kuawtli**	patlani	tepostototl	coyotl	*miak*
OISEAU :	***aigle***	*voler*	*avion*	*coyote*	*beaucoup*

Résultats (mots)



Guzman et al. 2025, Unigraphie MICAI

Tâche sémantique (phrases)

30 Références

5 Annotateurs

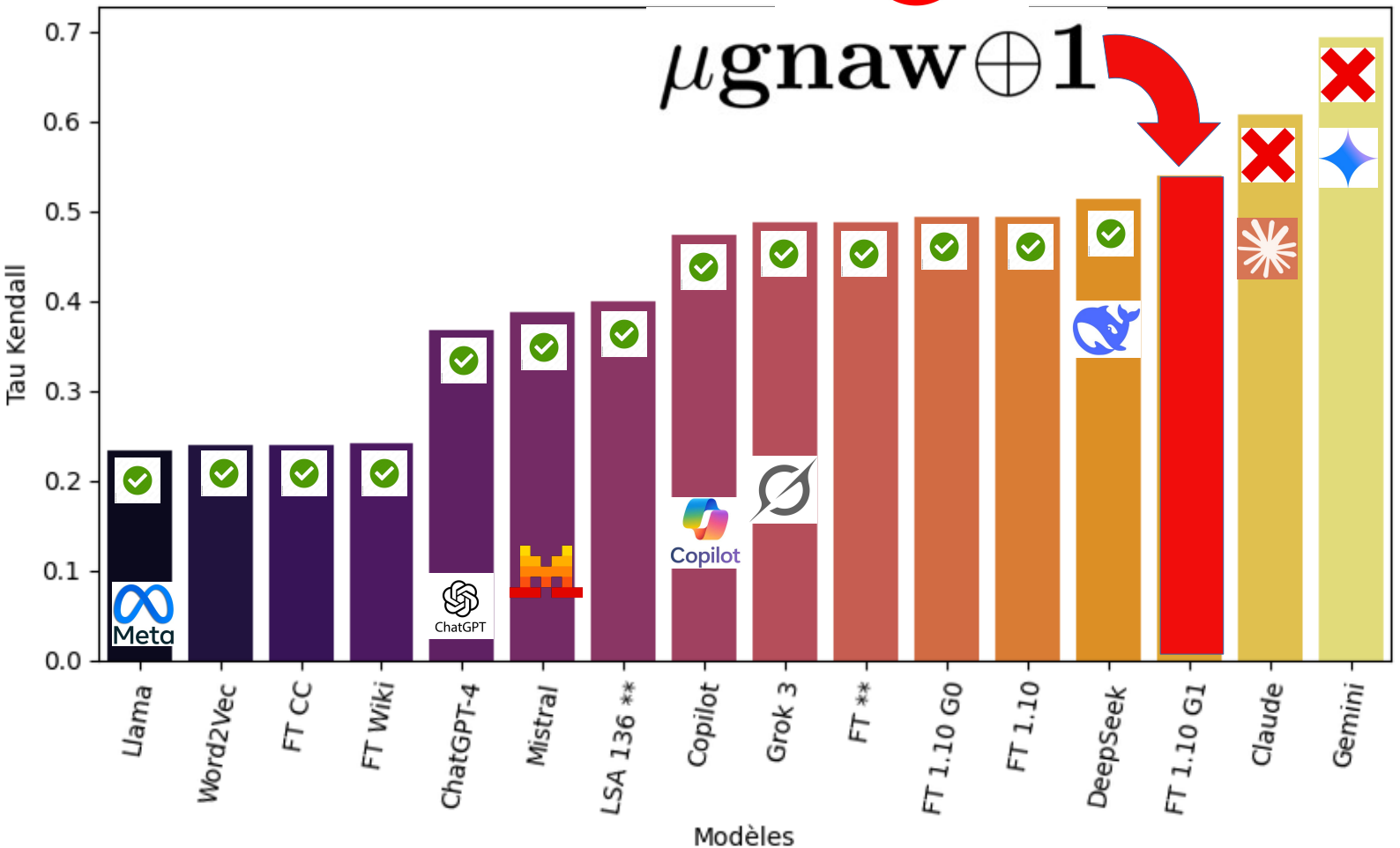
150 Candidates

REF$_2$: In posolli tlen nochipa nechpaktia motlalia ika tlaxkalli iwan epasotl / *La sopa que me gusta siempre se prepara con tortillas y epazote*

CANDIDATES:

1. Posolli ika tlaxkalli iwan epasotl nikneltoka ok achi welik ihkin amo ika kesoh / *I find the tortilla and epazote soup more delicious with cheese*
2. Epasotl kiwelilia posolli. / *The epazote gives a particular flavor to soups*
3. Epasotl ahwiyalli xiwitl tlen se kitekiwia ipan mexihkatlakualli / *Epazote is an aromatic plant used in Mexican food*
4. Moneki ma miak totonik posolli pampa ok achi techpaktia / *The soup should be very hot to enjoy it more*
5. Satepan kiawitl witz, tiyetokeh ihkin posolli / *After that downpour, we were left in the soup*

Résultats (phrases)



Guzman et al. 2025, Micrograммaires nawatl, soumission LREC'26

Conclusion : *embeddings…*

- **Explicables**: Nb **paramètres petit** et **connu** ; similarité **sémantique** calculée **intuitivement**
- **Stables**: les résultats **ne changent pas** entre chats (ce qui arrive souvent aux LLM)
- **Pas d'hallucinations**: phénomène fréquent et gênant affectant les LLM
- **Economiques** → **écologiques**

NAHU²: suite (sans conclusion)

- **Augmenter** le corpus (*Toujours! > +10M*)
  - OCR + Wiki + Audio + ReSoc + Livres sans © + Textes creation…
  - Texte artificiel → grammaires CFG
- **Evaluations** TALN (Sentiments, REN, Résumé automatique...)
- *https://demo-lia.univ-avignon.fr/pi-yalli*
  - Corpus annoté (meta-données)
  - Embeddings pre-entraînés
  - Transformateurs ▶ **BERTL**

¡Tlashokamati miak!

Merci beaucoup !

►Juan-Manuel Torres ● Martha-Lorena Avendaño ● Miguel Figueroa-Saavedra ● Graham Ranger ● Carlos González ● Patricia Velázquez ● Elvys Linhares ● Ligia Quintana ● Juan-José Guzmán ● Luis-Gil Moreno ● Jesús Vázquez◄

LABORATOIRE INFORMATIQUE D'AVIGNON
Université d'Avignon et des Pays de Vaucluse

I IDENTITÉ C CULTURELLE T TEXTES T THÉÂTRALITÉ

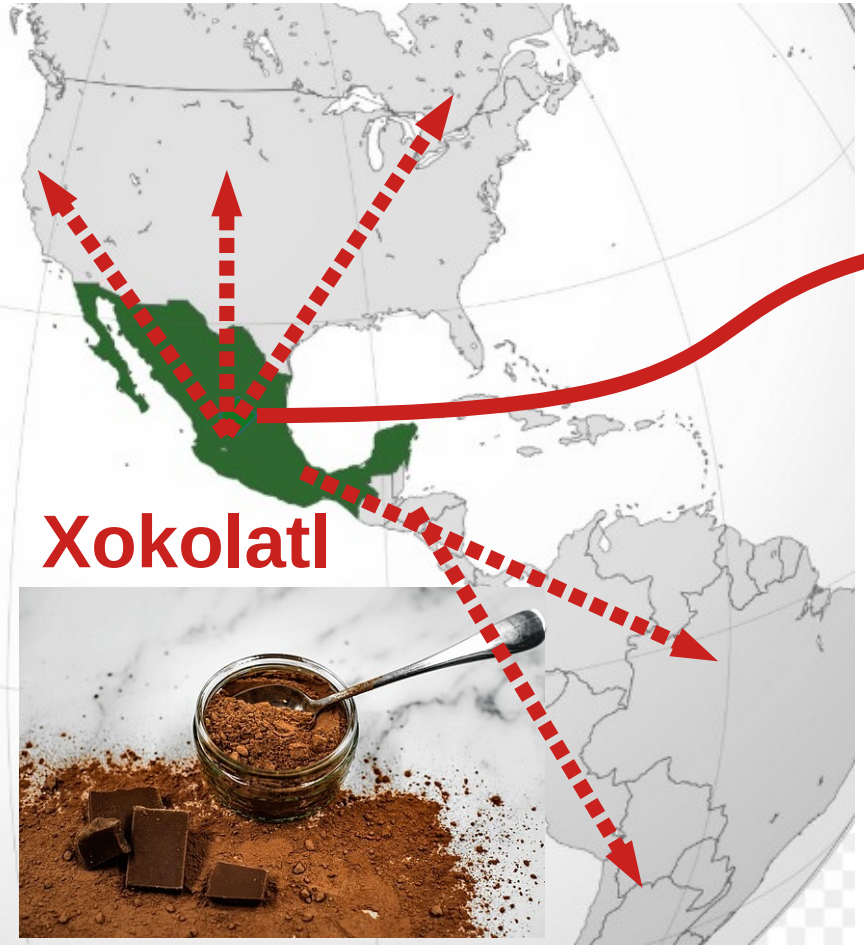
Mictlantecuhtli

Mictlan : Mikki + tlan = « Mort+lieu »

Mictlan+Tecuhtli = « Mictlan+Seigneur »

« Le Seigneur du domaine de la mort »

600 mots **nawatl** → **espagnol**



Xokolatl

Chocolate

Txokolate

Chocolat

Xocolata

Chokolade

Chocolade

Czekolada

Schokolade

Çikolata

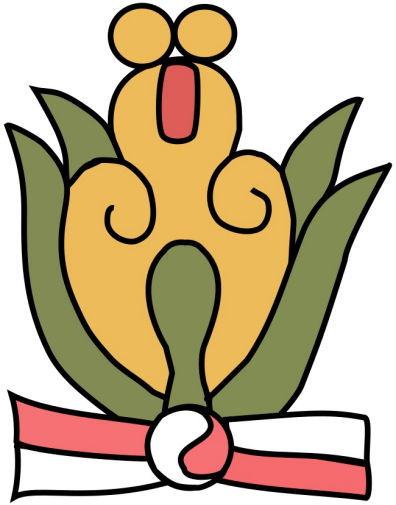
شكولاتة

Cioccolato

Shukulaato

Composition

- Huitzilopochtli : **Huitzil***in* + **opochtli** = *Colibri du sud?*
- Quetzalcoatl : **Quetzal***li* + **Koatl**
- Xochicuahuitl = **Xochi***tl* + **Kuawitl** = *Arbre fleurit*
- Axolotl = **A***tl* + **Xolotl** = Eau + *Monstre* = *Monstre aquatique*



Me(*tz*)+xi(*k*)+ko

- **Me**tztli = *Lune*
- **xi**ktli = c*entre*
- **ko** = *endroit*

Dans le centre de la Lune

π-langues

- **Ingredients**
  - **Corpus** représentatifs (taille, variantes linguistiques, sujets...)
- **Outils**
  - **Unificateur** de graphies
  - **Segmenteur** de termes (tokenisateurs)
  - **Représentations** (Modèles de langue, embeddings, transformateurs)
  - **Analyseur morphologique**
    - **Lemmatiseur / Stemmiseur (?)**
  - **Traducteur**

Mais il n'y a pas de corpus suffisants…