Nawatl Context-Free Grammars for Natural Language Processing

Juan-José Guzmán-Landa[1], **Juan-Manuel Torres-Moreno**[1,3], **Graham Ranger**[1], **Miguel Figueroa-Saavedra**[2], **Ligia Quintana Torres**[2], **Carlos-Emiliano González-Gallardo**[3,1], **Luis-Gil Moreno-Jiménez**[4], **Martha-Lorena Avendaño-Garrido**[2]

[1]LIA/ICTT Avignon Université, France [2]Universidad Veracruzana, Mexico
[3]LIFAT Université de Tours, France [4]Independent Researcher

{juan-jose.guzman-landa,juan-manuel.torres,graham.ranger}@univ-avignon.fr,
{migfig,lquintana,maravendano}@uv.mx, gonzalezgallardo@univ-tours.fr, lmoreno30470@gmail.com

Abstract

The aim of this article is to introduce Context-Free Grammars (CFG) for the Nawatl language. Nawatl is an Amerindian language of the π-language type, i.e. a language with few digital resources. For this reason the corpora available for the learning of Large Language Models (LLMs) are virtually non-existent, posing a significant challenge. The goal is to produce a substantial number of syntactically valid artificial Nawatl sentences and thereby to expand the corpora for the purpose of learning embeddings (static models or probability LLMs). For this objective, we introduce two new Nawatl CFGs and use them in generative mode. Thanks to these grammars, it is possible to expand the Nawatl corpus significantly and subsequently to use it to learn embeddings (such as FastText) and to evaluate their relevance in semantic similarity tasks. The results show an improvement compared to the results obtained using only the original corpus without artificial expansion, and also demonstrate that economic embeddings often perform better than some LLMs.

Keywords: Nawatl, Symbolic NLP, Corpus, Context-Free Grammars, Semantic similarity

1. Introduction

Nawatl or Nahuatl is one of the Indigenous Languages of Mexico and the most widely spoken national Indigenous languages, with approximately 1.65 M speakers (INEGI, 2020). The language is marked by substantial dialectal diversity, comprising 29 recognized variants distributed across four main regions: the Western, Central, Eastern, and Huasteca zones (Ethnologue, 2025[1]; Lastra de Suárez, 1986). This linguistic diversity presents challenges for the use of textual corpora in educational, communicative, and digital contexts, as it entails a significant variation in orthography and lexical choice (Zimmermann, 2019; Olko and Sullivan, 2016; Hansen, 2024; Figueroa-Saavedra, 2024).

Although the publication of digital material in Nawatl has increased, the notable dispersion and variety of such resources has prevented them from gaining visibility on social media and from being clearly identified and accessed in repositories. This is nonetheless not an obstacle to the gradual increase in their presence and literate use, even if the digital linguistic resources available, which are essential for the current revitalization of the language, remain limited (Pugh et al., 2025). Therefore, a serious problem is the scarcity of computational resources for this language and, in particular, corpora available for machine learning. In this context, could symbolic formal grammars mitigate this major problem?

Our approach to tackle this problem proposes the generation of artificial corpora that respect the structure of Nawatl language using formal grammars. Our goal in proposing the grammar is not to model the complexity of the Nawatl language fully, but rather to generate syntactically valid sentences. These sentences can serve as a basis for the creation of large-scale synthetic corpora. Such corpora may be leveraged to enhance the training of both static and dynamic context Large Language Models (LLMs) for π-languages -i.e. languages with limited digital resources (Berment, 2004; Abdillahi et al., 2006). Specifically, we aim to expand a Nawatl corpus, as *Axolotl* (Gutierrez-Vasques et al., 2016), a corpus of documents in bilingual Spanish/Nawatl versions[2], the Ihquin tlahtouah in Tetelahtzincocah corpus (Pugh et al., 2025) or π-YALLI[3] (Guzmán-Landa et al., 2025a) a corpus has previously been employed in word-level semantic similarity tasks (Torres-Moreno et al., 2024). The size of the vocabulary has a direct impact during training: a larger vocabulary is expected to positively affect the performance of the model (Tunstall et al., 2022; Goyal et al., 2018).

[1]https://www.ethnologue.com

[2]The *Axolotl* corpus can be retrieved from: http://www.corpus.unam.mx/axolotl

[3]https://demo-lia.univ-avignon.fr/pi-yalli

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3333–3342
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

The structure of the paper is as follows: Section 2 provides an overview of the Nawatl language and its grammatical features and introduces Context-Free Grammars. Section 3 presents our proposed CFG micro-grammars for Nawatl. Section 4 describes the extended corpus, π-yall-IA. Section 5 reports on experiments using the micro-grammars applied in a semantic similarity task. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Nawatl and Formal Grammars

2.1. Nawatl grammar at glance

Nawatl, an Uto-Aztecan language, is an agglutinative and polysynthetic language that composes words by joining various morphemes to a verbal or nominal root, thereby constructing meaning. Nawatl sentences have a basic Verb-Subject-Object (**VSO**) syntax, but allow for structural flexibility that responds to speakers' needs. Thus, orders such as **VO, VS, VOS**, and, less frequently, SV, SVO, and SOV can also be found (see Table 1).

Syntactic relationships between words and clauses are established through the valency of the verb and the use of connectors (particles). These connectors can also be formed through groupings that establish nuances of meaning, in addition to discourse connectors. Some of these words can form "one-word sentences", since their morphology includes the subject and predicate, as well as information about the actants and modal, directional, and relational elements (see Launey, 1978; Flores Nájera, 2019; Sasaki, 2022).

Structure	Example	Translation
VSO	*Kitta tlakatl kalli*	Sees (a) man (a) house
VO	*Kitta kalli*	(He/she) sees (a) house
VS	*Kitta tlakatl*	Sees (him/her/it) (a) man
VOS	*Kitta kalli tlakatl*	Sees (a) house (a) man
SV	*Tlakatl kitta*	(A) man sees (him/her/it)
SVO	*Tlakatl kitta kalli*	(A) man sees (a) house
SOV	*Tlakatl kalli kitta*	(A) man (a) house sees

Table 1: Syntactic constructions in Nawatl language. More frequent structures (in relation to the verbs) are indicated in bold.

To our knowledge, there is no formal context-free grammar for Nawatl. In the following section, we will present Context-Free Grammars (CFG), and also introduce the first two CFGs for this language.

2.2. Context-Free Grammars

A Context-Free Grammar is a formal grammar used to describe the syntax of formal languages, especially in the syntactic analysis of programming languages and natural languages (Hopcroft et al., 2006). Formally, a CFG is a quadruple: $G = (V, \Sigma, R, S)$ where:

- 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: $A \rightarrow \alpha$ 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.

A grammar is called context-free because the rules are applied without considering the context in which the non-terminal symbol appears.

3. Two new Nawatl Context-Free Grammars

This section introduces two new context-free micro-grammars for Nawatl. These grammars are considered micro versions mainly because they both avoid the use of recursive production rules. Similarly, only some grammatical persons are included, and verbs are limited to the singular form and the present tense.

3.1. μGNAW$\oplus$0: a first Nawatl CFG

This approach is inspired by the grammatical frameworks developed for Indo-European languages. Given that Nawatl belongs to Indigenous American language families -linguistically distant from Indo-European families- this model results in a limited and reductive representation of the language.

This grammar develops a classical model consisting of two types of phrases: the noun phrase (N) and the verb phrase (V). Both types can include additional elements (particles, nouns, and prefixes) that translate into our grammatical categories as temporal adverbs and quantifiers (ADV_T, ADV_Q), adjectival nouns (ADJ), personal pronouns (PP) and person markers (PV), possessive markers (POS) and negation (NEG), see Figure 1.

Given that Nawatl belongs to the families of Indigenous American languages, which are linguistically distant from the Indo-European family, such a model results in a limited and reductive representation of the language. Nevertheless, the objective here is not to provide an accurate model of real Nawatl, but rather to illustrate that the conventional structure: P$\rightarrow$ N (noun phrase) V (verb phrase), where V may also contain N, is insufficient to capture the deeper syntactic organization of Nawatl, which is centered around the verb. Although derived from the grammars of families of well-resourced languages, this preliminary version

```
P → ADV_T (N|V)
N → ADJ (ART_|POS)⊕n
V → N NEG PV_3⊕v ADV_Q
V → PP_i NEG PV_j⊕v ADV_Q ; i , j =1 ,2 ,3; i=j

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

n → siwatl|miston|elotl|xokotl|tochin|
    yolkatl|nakatl|···
  # woman|cat|corn|fruit|rabbit|animal|meat|···
v → nehnemi|kwa|kaki|···
  # to walk|to eat|to listen|···
⊕=concatenation ∅=null _=space
```

Figure 1: CFG Nawatl μGNAW⊕0.

nonetheless enables the generation of some basic Nawatl structures.

We present here a few examples of the production rules of the micro-grammar, in generative form. For instance, the noun phrase N can generate:

- ADJ ART n: Yehyektsin ni/ne siwatl
  # Beautiful [is] this/that woman
- ADJ POS n: Tomawak motoch
  # Fat [is] your rabbit

Similarly, the verb phrase V could yield productions such as the following:

- POS_2⊕ n NEG PV_3⊕ v: momiston amo _nehnemi *# your_cat it doesn't walk*
- PP_1 NEG PV_1⊕ v ADV_Q n: na amo nikkwa miyak xokotl *# It's me (who) I don't eat a lot [of] fruit*
- ADV_T ADJ (ART |POS)⊕ n NEG PV_3⊕ v ADV_Q n: axan tomawak ni /no/toch(in) amo _(tla)kwa tlawel elotl *# Now fat the/my rabbit doesn't eat too much corn*

3.2. μGNAW⊕1: a realistic Nawatl CFG

The new CFG Nawatl grammar is based on the prototypical grammatical structures characteristic of the VSO word order (and derivatives), as well as on the less frequent SVO patterns described in Section 2.

The main objective is to accurately model the different types of verb phrases V, which are frequently used, along with the relatively less common noun phrases S. These phrases can incorporate elements that we will call *markers*. Some of these markers have agglutinative characteristics, and others are lexical particles that can indicate place, time, and intensity, playing a role similar to that of "adjectives" in Indo-European languages. We have therefore defined the following: person marker (MV), object marker (MO), possessive marker (POS), temporal marker (MT), quantity marker for nouns (MCS), intensity marker for verbs (MIV), and finally the place marker (ML). In addition to the above, there are terminal nodes that represent negations (NEG), adjectives (ADJ), nouns (n), and verbs (v).

To achieve a better approximation of Nawatl grammar, the following modifications have been introduced:

- The numeral **se** (one) and the particle **ni / in** (the) have been omitted;
- The personal markers **na / neh** (I/my), **ta / teh** (you), and **ya / yeh** (he/she/it) have been excluded;
- Quantifier and temporal adverbs have been reclassified as MCS, MIV and MT;
- Place markers ML have been introduced;
- Nouns (n) now have two forms, one that allows them to be possessed (with a POS tag) and one that does not.

In this micro-grammar (see Figure 2), the verb assumes a central and predominant role in Nawatl syntactic constructions. Thus, the formulation of this new grammar is based on structural patterns observed in the Nawatl spoken by Nahua speakers.

```
P → VSO | V⃗O | V⃖S | VOS | SV | SOV | SVO
V → NEG MT MIV MV_i⊕MO_j⊕v ; i , j =1 ,2 ,3; i≠j
S → MCS ADJ POS⊕n
O → ADJ POS⊕n ML

ADJ → weyi|istak|∅      # big|white|∅
POS → no|mo|i|∅         # my|your|his|∅
NEG → amo|∅             # no|∅
MT → aman|cemicac|∅     # now|forever|∅
MIV → miyak|nochi|∅     # a lot|all|∅
MCS → san|miakpa|∅      # only|often|∅
ML → nikan|nepa|∅       # up|there|∅
MV_i →∅                 # he,she
MO_j →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|···
⊕ = concatenation ∅ = null
```

Figure 2: CFG Nawatl μGNAW⊕1.

The symbols $\overrightarrow{VO}$ and $\overleftarrow{VS}$ respectively indicate the reading direction of the structure to minimize ambiguities between the subject and the object. For example, the sentence $\overleftarrow{VS}$: *Kitta tlakatl*: A man sees (him, her, it), unlike the sentence $\overrightarrow{VO}$: *Kitta kalli*: (He/she) sees a house.

Unlike micro-grammar μGNAW$\oplus$ 0, the micro-grammar μGNAW$\oplus$1 can establish more syntactic relationships between words in their text normal composition. To do this, we impose the following restrictions:

- Exclusive use of base 1 verbs.[4]
- Exclusive use of transitive verbs and verb conjugation and valency in the 3rd person singular.
- Adjectives are part of the S and O structures, that is, they are positioned before POS.
- There are two markers of intensity or quantity: MIV is used to intensify the verb, and MCS to quantify the noun.

All of this allows us to add more features closer to real Nawatl within the micro-grammar μGNAW$\oplus$1.

On the other hand, rhetorical connectors CR (and, but, more, however, etc.) add greater variability and richness to the generated sentences, and could be incorporated into the micro-grammar as follows: P $\rightarrow$ P CR P. However, due to its recursive nature, this rule will not be included in both grammars in their generative phase, but will be included as post-processing (see 4.2) in grammar μGNAW$\oplus 1$ with a variable number of CR, in order to increase diversity and produce a more realistic textual output: P $\rightarrow$ P CR P [[CR P] CR P ...].

4. An augmented artificial corpus

The grammars μGNAW$\oplus(\bullet)$ are capable of generating a large number $\mathfrak{F}^{0,1}$ of grammatical productions. However, although the value is large, the sentences produced represent only a tiny fraction of the number that could be produced using recursive grammars. Such recursive grammars, however, fall outside the scope of this article.

Nevertheless, the number of sentences remains sufficient to enable the artificial enrichment of the π-YALLI corpus. To achieve this, we must establish certain restrictions in order to produce not only grammatically correct but also semantically acceptable sentences. Indeed, we prefer not to accept sentences such as "*The big corn-cob eats a lot of rabbit*" or "*The fruit dreams too much*" because, although they belong to μGNAW$\oplus(\bullet)$, they are not semantically realistic. We must therefore implement filters.

[4]Verbs in Nawatl can have three bases: base 1 represents all verbs in the present tense, base 2 represents verbs in the past tense, and finally, base 3 represents verbs in the hypothetical future tense.

4.1. Semantic Filters

The artificial corpus derived from the grammars presented should contain sentences that can be effectively used for training static LLMs. In order to retain only sentences that are both grammatically correct and semantically acceptable, we propose to introduce a filter based on the association between verbs and animate/inanimate nouns, and no repeated nouns in the same sentence.

The animate/inanimate filter enables the elimination, at low computational cost, of a significant subset of sentences that lack reasonable semantics. The concrete implementation of the filter is as follows: a sentence is constructed using only nouns associated with verbs of the same type (n animate/v animate or n inanimate/v inanimate), avoiding the mixing of animate/inanimate ones. By example:

Nouns:

- *nantzin* (mother): **animate**
- *tiotzin* (god): **animate / inanimate**
- *tatzin* (father): **animate**
- *tlahtolli* (speech): **inanimate**
- *mapachin* (raccoon): **animate**
- *kuawtli* (eagle): **animate**
- *momachtiani* (student): **animate**

Verbs:

- *mawisowa* (admire): **animate**
- *neki* (want): **animate**
- *pia* (to have): **animate / inanimate**
- *itta* (to see): **animate**
- *chihua* (to do): **animate**
- *toka* (bury): **animate / inanimate**

The rule states that if the labels match, the sentence can be generated. That is, the nouns father (*tatzin*), mother (*nantzin*) or raccoon (*mapachin*) can be generated with the verbs want (*neki*) or see (*itta*), but this is not the case for the noun speech (*tlahtolli*).

In summary, based on micro-grammars μGNAW$\oplus$(0,1), sets of $\mathfrak{F}^{0,1}$ sentences are generated. These sets, once properly filtered, contain respectively $\mathfrak{F}^{0,1}{*}$ sentences, where $\mathfrak{F}^{0,1} > \mathfrak{F}^{0,1}{*}$. This procedure allows us to retain only a diverse set of grammatically and semantically acceptable sentences, which constitute an artificial and usable corpus of sentences in Nawatl. We present some examples for our μGNAW$\oplus$0:

- *Aman weyi ni ichpochtli amo toka miakeh* / Now, lady doesn't bury several ones.
- *Aman weyi ni ichpochtli amo ahsikamati nochi* / Now big lady doesn't understand everything.

- *Axkan weyi mokoltzin amo maka miyak* / Now grandfather doesn't give much.

μGNAW$\oplus$0 generates $\mathfrak{F}^0 \approx 1 \times 10^6$ unfiltered sentences, or $\mathfrak{F}^0* = 807,093$ filtered sentences. On the other hand, μGNAW$\oplus$1 generates $\mathfrak{F}^1 \approx 4.72 \times 10^{12}$ unfiltered phrases, which is a very large number. Even applying the semantic filter is unrealistic. For this reason, we use another sentence filtering strategy based on symbolic tags that limit this combinatorial explosion (see Section 4.2). The number of sentences produced by grammars can be calculated using the values in tables 2 and 3 for terminal nodes:

n	v	ADV_Q	POS	ART	$\ddots$
26	16	5	3	3	
$\ddots$	ADV_T	ADJ	PP	NEG	$\mathfrak{F}^0$
	7	3	3	3	1×10^6

Table 2: Number of sentences for μGNAW$\oplus$0.

n	v	MT	MIV	MV	MO	$\ddots$
42	27	11	4	1	1	
$\ddots$	MCS	ADJ	POS	NEG	ML	$\mathfrak{F}^1$
	4	10	4	11	8	4.72×10^{12}

Table 3: Number of **VSO** sentences for μGNAW$\oplus$1.

We decided to implement the micro-grammars μgnaw$\oplus(\bullet)$ in Prolog, an expressive language well suited to the generation of symbolic Context-Free Grammars. In addition, Prolog is fast and manages memory efficiently.

4.2. Post-processing of μGNAW$\oplus$1

This section details the processes of normalizing sentences (cleaning, capitalization, etc.), introducing symbolic tags to control the combinatorial explosion of the grammar and to promote diversity in its realizations.

It also describes the processes of segmentation into paragraphs and the stochastic introduction of rhetorical connectors with the aim of producing realistic artificial text for our purposes. These stages are illustrated in Figure 3 and described below.

Symbolic tagging. This labeling consists of replacing the values of the nodes NEG, MCS, MIV, MT, ML, and ADJ with their respective tags [NEG], [MCS], [MIV], [MT], [ML] and [ADJ]. These tags are then replaced with values from the terminal nodes, respecting the distribution probability in the authentic corpus. This gives the artificial corpus more realistic characteristics.

Possessive management POS. Phrase generation transforms nouns when they are possessed by a POS particle. For example, if *ichpochtli* (**lady**) is used, and you want to say: **my lady**, the absolutive *tli* must be removed. Thus, **my lady** is written: *ni* (**my**) + *ichpoch(tli)* = *noichpoch*.

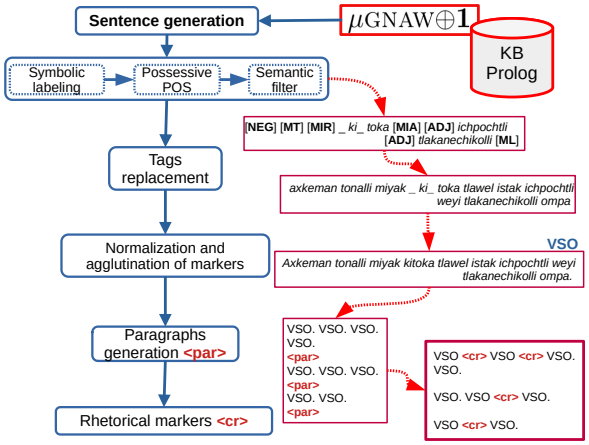

Figure 3: Post-processing for μGNAW$\oplus$1.

Paragraph generation. The paragraph tagger applied to the corpus π-YALLI generates 8,767 paragraphs from its 364,233 sentences[5]. On the other hand, μGNAW$\oplus$1 produces 1'956,750 sentences in a single long paragraph. This paragraph was segmented respecting the proportion of paragraphs found in π-YALLI. This yields 41,220 artificial "paragraphs" that simulate the number of authentic paragraphs.

Introduction of rhetorical connectors. A paragraph can consist of one or more VSO structures. Each paragraph is analyzed and, given the probability of a rhetorical connector, the "." separating two VSO structures is replaced by a rhetorical connector.

Text normalization. This process requires the following steps: removing repeated phrases, concatenating words, normalizing spaces, capitalizing the first letter, and inserting a period.

Table 4 shows the basic statistics for both grammars, used in generative mode.

Grammar	Tokens	Sentences	Paragraphs
μGNAW$\oplus$0	$\approx$ 4.6M	$\approx$ 807K	714
μGNAW$\oplus$1	$\approx$ 6M	$\approx$ 1.9M	41 220

Table 4: Statistics on artificial corpora μGNAW$\oplus(\bullet)$. Central Nawatl variant.

We present some examples generated by μGNAW$\oplus$1:

[5]Computed from a normalized corpus without empty lines, duplicated spaces, commented lines, etc

1 *Kimachtia itoto tonatih. Kimachtia itoto sentilistli. Kimachtia itoto momachtiani. Kimachtia itoto temachtiani. Tlen Kimachtia itoto posoli.*

2 *Kitoka mokoltsin noxochih. Kitoka mokoltsin nosiwah. Kitoka mokoltsin notatsin. [...] Kitoka mokoltsin itlaka.*

4.3. Merging artificial sentences and the Nawatl π-YALLI corpus

To our knowledge, no attempt to expand the corpus in the Nawatl language has been made to date. Indeed, there are studies that point in this direction, particularly in spoken language (Bartelds et al., 2023). Our emergent idea is the combination of an "authentic" corpus such as π-YALLI with an artificially generated corpus. We think that this union may be beneficial for the training of Language Models (whether static or LLM). In particular, we have focused on static LM training. After combining the two corpora, the new corpus goes through an elementary process of unification of spellings proposed by (Guzmán-Landa et al., 2025b), in order to standardize the use of characters of the different variants of Nawatl[6].

We now present some characteristics of the –authentic– π-YALLI Nawatl corpus (Guzmán-Landa et al., 2025a), that we have used for our semantic similarity experiments.[7] The available corpus is heterogeneous in terms of categories and linguistic variants in Nawatl (see Table 5), and contains a relatively small number of tokens ($\approx$ 6M) and sentences ($\approx$ 428K)[8].

This makes it useful for training classical language models (vector models, TF.IDF, etc.) or static models (Word2Vec, FastText, or Glove embeddings), but unsuitable for training contextualized language models (using BERT-type transformers). Indeed, it has been reported that contextual LLMs require in the order of 10M-100M tokens to achieve stable embeddings (Micheli et al., 2020; Hoffmann et al., 2022).

For this reason, we decided to unify the π-YALLI corpus with the filtered artificial sentences generated by the grammar. The enlarged corpus was then employed to train FastText static embeddings with the objective of improving performance on the task of semantic similarity between sentences.

[6]For example, changing c for k, hu for w, the elimination of double consonants, the use of lowercase letters, avoiding accents, etc.

[7]The π-yalli corpus may be found on the website: https://demo-lia.univ-avignon.fr/pi-yalli

[8]Values calculated on the raw text corpus.

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**	**$\approx$ 6 629 000**	**$\approx$ 428 000**

Table 5: Statistics on the π-YALLI v1.10 corpus. AGR: Agriculture; COS: Cosmovision; ECO: Economy; EDU: Education; HIS: History; LEG: Legal documents; LIN: Linguistics; LIT: Literature; MED: Medicine; MUS: Music; PHR: General sentences; POE: Poetry; REL: Religion; TEC: Science and Technology; WIK: Wikipedia.

5. A sentence-level semantic similarity task

In this section, we present a protocol to evaluate our approach. Semantic similarity is a standard task in Natural Language Processing (NLP), involving the evaluation of various models (statistical models, neural networks, etc.) through standardized evaluation protocols (Francis-Landau et al., 2016). In particular, we focus our attention on the task of semantic similarity between a reference sentence and a set of candidate sentences. This results in an ordered list of candidates that can be compared against a ranking created by a human. This is the evaluation protocol used in (Guzmán-Landa et al., 2025a), which we adopt here by training embeddings to assess the impact of learning based on both existing and artificially generated corpora.

Figure 4 illustrates the complete experimental pipeline we designed, from the construction of the corpus (including authentic documents or generated from the formal grammar) to the evaluation of the semantic task via a reference ranking produced by human annotators, compared to the ranking generated by FastText.

We did not expect to achieve high performance in the semantic similarity task using only the artificial corpora generated by the micro-grammars μGNAW$\oplus(\bullet)$. Rather, our goal is to enrich the π-YALLI corpus with new artificial datasets containing frequently used grammatical structures. Another of our hypotheses is that the corpus μGNAW$\oplus$1 -which better models the Nawatl grammar- should contribute to increasing the performance of embedding-

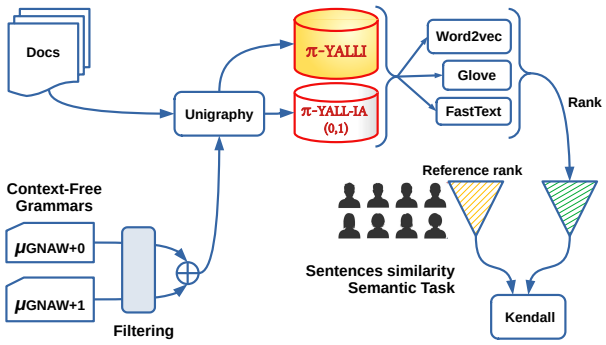

Figure 4: The Sentences Semantic Similarity task.

based models.

The extended corpora π-YALL-IA$\oplus(\bullet)$ were used to train the FastText algorithm (Bojanowski et al., 2017) from scratch. The resulting embeddings were subsequently used in the task of semantic similarity between sentences.

A total of $R = 30$ references were written in Spanish by an annotator. These R sentences were divided into 6 blocks of 5 sentences each. Each block was assigned to an annotator who drafted $C = 5$ candidate sentences with varying degrees of semantic similarity relative to the original sentence. A bilingual Nawatl speaker (Spanish-Nawatl) translated into the Central Nawatl variant all sentences (references and candidates). The basic statistics for this task are as follows in Table 6.

	Sentences	Tokens	Types	Tokens per Sentence
References	30	246	189	8.20
Candidates	150	1026	599	6.84

Table 6: Statistics of the semantic task.

Here we show one example of the semantic similarity classification task. Candidate phrases are sorted from highest to lowest semantic similarity with respect to their reference.

REFERENCE #3:
Tonatih kipalewia moskaltiah kilitl iwan xokokuawitl / The sun serves to make green plants and fruit trees grow.
CANDIDATES:

1. *Moneki xiwimeh kiseliah tlawili, atl pampa moskaltiskeh* / Vegetation needs to receive light and water to grow.
2. *Kuawmeh kinmanawiah xiwimeh iwikpa tonalmitl* / Trees protect plants from the sun's rays.
3. *Amo nechpaktia nikilikua ipanpa ok achi nechpaktia tlakilotl* / I don't like eating vegetables because I prefer fruit.
4. *Tonaltlawili ekolohika ipanpa xoxowik tlawili, no ihkin atlawili* / Solar energy is environmentally friendly because it is a green energy source, like water energy.
5. *Ihkuak se kiselia miak tonalmitl weli kinextia kualokatl* / Receiving too much sunlight can cause cancer.

The reference ranking R_R is produced by humans, and the goal of the task is to measure how closely the rankings R_M generated by various language models approximate the reference ranking R_R. The measure of similarity between rankings would be estimated using Kendall's $\tau(R_R, R_M)$, a nonparametric measure of correlation that evaluates the ordinal association between two variables, i.e., the degree of agreement between two rankings x, y (with tie correction):

$$\tau = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}} \quad (1)$$

where C is the number of matching pairs, D the number of discordant pairs, T_x the number of tied pairs in x and T_y the number of tied pairs in y (Kendall, 1938).

5.1. Static embeddings

While it is true that transformers have demonstrated superiority in NLP tasks that use computationally rich languages, the situation changes when processing π-languages. Static embeddings can be generated from scratch, are quick to produce, and can be used with small corpora containing a certain number of words that appear rarely. Examples of these algorithms are Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017), and Glove (Pennington et al., 2014).

We found three sets of Nawatl pre-trained embeddings available: FastText pre-trained on Common Crawl[9]; or Wikipedia in Nawatl[10] and Word2Vec pre-trained[11]. We decided to compare FastText trained on our expanded corpora with these sets of pre-trained embeddings. FastText's implementation[12] was trained using the following hyperparameters: Iterations=20; Window=5; Mode=Skip-Gram; Dimensions=300. (Mikolov and Zweig, 2012) discovered that Skip-Gram works well with small corpora and can better represent less frequent words. Then, we decided to use this implementation with the Nawatl texts.

[9] https://commoncrawl.org/

[10] FastText has been trained on 157 languages worldwide. See the website: https://fasttext.cc/docs/en/crawl-vectors.html

[11] See website: https://sparknlp.org/2022/03/16/w2v_cc_300d_nah_3_0.html

[12] From: https://radimrehurek.com/gensim/models/fasttext.html

5.2. The LLMs employed

In this paper, we tested the following five LLMs models via APIs:[13] ChatGPT-4 mini API; Gemini-2.5-flash-preview-05-20 API[14]; DeepSeek-V3-0324 API[15]; Llama-3.1-70B-Instruct API[16] and Mistral-large-latest API.[17] In an interactive user mode, we employed three well-known LLM models: Copilot (https://www.microsoft365.com), Grok 3 (https://grok.com) and Claude 3.7 (https://claude.ai).

The prompt used was the following:[18]

> *Given the Nawatl sentence "[reference.]" rank the following five sentences semantically, from the closest to the furthest in meaning from the original sentence: "[candidate$_1$.]"; "[candidate$_2$.]"; "[candidate$_3$.]"; "[candidate$_4$.]"; "[candidate$_5$.]". Do not give ranking of other sentences than those provided.*

In our evaluation, we used a cross-validation protocol (Leave-three-out) (Moss et al., 2018), which avoids possible overfitting biases.

5.3. Results and Discussion

Figure 5 summarizes all our results. As shown in this Table, the best result of FastText was obtained using the filtered π-YALL-IA$_{v1.10}\oplus$1 corpus, which yielded a Kendall's τ = **0.540**. This places the algorithm FastText model trained with an expanded corpus in 3rd place in the ranking, only behind the LLMs Gemini 2.5 and Claude 3.7. As expected, the grammar μGNAW$\oplus$0 was disappointing. But this is good news, because it shows that grammars closer to Nawatl VSO structures are very relevant for expanding existing corpora.

6. Conclusions and Future works

The generative use of micro-grammars μGNAW$\oplus$(0,1) allows for the creation of a large number $\mathfrak{F}^{0,1}$ of sentences, which can be considered as a corpus -artificial and very extensive- but which presents redundancy and poor semantics. A semantic filtering process was applied in order to reduce the universe to $\mathfrak{F}^{0,1}*$ sentences, decrease redundancy, and achieve slightly more plausible semantics. Subsequently, the artificial corpus was concatenated with the authentic corpus. π-YALL-IA$\oplus$1, a corpus resulting from the union of μGNAW$\oplus$1 (closer to the structures of the Nawatl language) and π-YALLI, enabled efficient learning using a representation of static embeddings. Using π-YALL-IA$\oplus$1, the FastText algorithm improved its performance in the task of evaluating semantic similarity between sentences, increasing the Kendall τ_{MAX} from 0.493 to 0.540 (+9% increase).

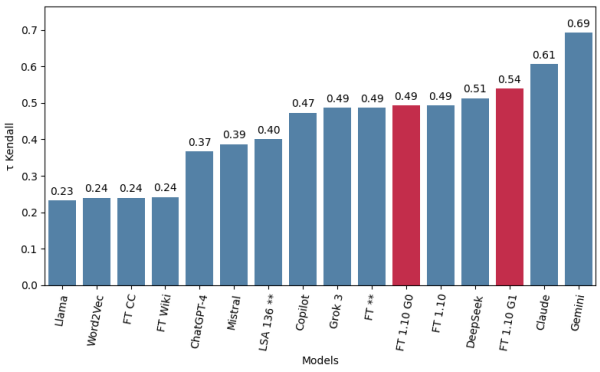

Figure 5: Kendall's τ: LLMs, FastText/Common Crawl-Wikipedia (FT CC / FT Wiki) and Word2Vec vs FastText/π-YALL-IA Skip-gram in 300 dimensions, without empty words: *iwan, n, tlen* & *ipan*. Maximum values obtained in *Leave-3-out* evaluation are shown (red bars indicate the Kendall's τ using our CFGs). ** LSA results come from (Guzmán-Landa et al., 2025b)

It should be noted that in this task, we also compared FastText and Word2Vec embeddings pre-trained on other corpora, and their results are disappointing ($\tau \approx 0.242$) compared to FastText trained on π-YALLI. Using a formal Nawatl micro-grammar in generative mode seems to favor learning FastText algorithms because the embeddings better capture the structure of the Nawatl language.

In future work on the grammar μGNAW$\oplus$1, we will seek to increase the number of elements n and v, include additional grammatical persons, base 2 and 3 verbs, plurals, more rhetorical connectors, markers, and improve the semantic filter (to avoid combinatorial explosion effects). Another idea to explore is the introduction of Probabilistic Context-Free Grammars (PCFG), where nodes can have associated probabilities. This type of grammar[19]

[13] The API accesses to LLMs were performed on 09/06/2025 from 16h-20h GMT. The interactive ones (Copilot, Grok, and Claude) were performed on June 9-10/2025.

[14] https://deepmind.google/models/gemini/flash/

[15] https://api-docs.deepseek.com/news/news250325

[16] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[17] https://mistral.ai/news/mistral-large

[18] The prompt was written in French: « *À partir de la phrase nawatl: "[référence.]" triez par ordre sémantique, de la plus proche à la plus lointaine, les cinq phrases suivantes: "[candidate$_1$.]"; "[candidate$_2$.]"; "[candidate$_3$.]"; "[candidate$_4$.]"; "[candidate$_5$.]". Ne donnez pas des rankings de phrases autres que celles proposées.* »

[19] https://en.wikipedia.org/wiki/Probabilistic_context-free_grammar

(Smith and Johnson, 2007) is particularly useful when filtering semantic/non-semantic sentences using an external classifier. This classifier begins with manual labelling and then produces automatic labelling of a subset of phrases. This classification could be used to modify the parameters (in this case, the probabilities) of the generative grammar.

7. Limitations

Our findings indicate that traditional statistical techniques outperform more sophisticated approaches. Although these results are highly promising, we acknowledge that additional experimentation with larger and more representative datasets is necessary, particularly when dealing with π-languages that have limited computational resources.

8. Ethical considerations

The use of LLMs under this study focuses on comparing semantic similarity performance against classic statistical methods. Even if this is a low-risk task, we emphasize that LLM-generated outputs should always be carefully validated.

9. Acknowledgments

This research work has been financed by the Agorantic NAHU² project and the Intermedius PhD Grant, and partially supported by the Laboratoire Informatique d'Avignon (LIA), from Avignon Université (France).

10. Bibliographical References

Nimaan Abdillahi, Pascal Nocera, and Juan Manuel Torres. 2006. Boites a outils TAL pour les langues peu informatisées : Le cas du Somali. In *Journées d'Analyses des Données Textuelles*, Besançon, France.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. ACL.

Vincent Berment. 2004. *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"*. Ph.D. thesis, Université Joseph-Fourier - Grenoble I.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Miguel Figueroa-Saavedra. 2024. *Amapowalistli iwan tlahkuilolewalistli*. Universidad Veracruzana, Mexico.

Lucero Flores Nájera. 2019. *La gramática de la clausula simple en el náhuatl de Tlaxcala*. Ph.D. thesis, CIESAS.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1256–1261, San Diego, California. ACL.

Palash Goyal, Sumit Pandey, and Karan Jain. 2018. *Deep Learning for Natural Language Processing*. Springer.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *10th LREC'16*, pages 4210–4214.

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Graham Ranger, Martha Lorena Garrido-Avendaño, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Carlos-Emiliano González-Gallardo, Elvys Linhares-Pontes, Patricia Velázquez-Morales, and Luis-Gil Moreno Jiménez. 2025a. π-yalli: un nouveau corpus pour le nahuatl / *Yankuik nawatlahtolkorpus pampa tlahtolmachiotl*. In *TALN Marseille*, pages 802–816. ATALA.

Juan-José Guzmán-Landa, Jesús Vázquez-Osorio, Juan-Manuel Torres-Moreno, Graham Ranger, Martha Lorena Garrido-Avendaño, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Patricia Velázquez Morales, and Gerardo Sierra-Martínez. 2025b. A symbolic algorithm for the unification of nawatl word spellings. In *MICAI'25*, page 12p. SMIA.

Magnus Pharao Hansen. 2024. *Nahuatl Nations: Language Revitalization and Semiotic Sovereignty in Indigenous Mexico*. Oxford University Press.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy,

Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *arXiv abs/2203.15556*.

John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2006. *Introduction to Automata Theory, Languages, and Computation*, 3rd edition. Pearson.

INEGI. 2020. Censo de población y vivienda 2020. https://www.inegi.org.mx/rnm/index.php/catalog/632/study-description.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Yolanda Lastra de Suárez. 1986. *Las áreas dialectales del náhuatl moderno*. UNAM, Instituto de Investigaciones Antropológicas, Mexico.

Michel Launey. 1978. Introduction à la langue et à la littérature aztèques. *Paris, L'Harmattan*, 1.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Conference on Empirical Methods in Natural Language Processing*, pages 7853–7858. ACL.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS - Vol 2*, NIPS, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 234–239, New York, NY, USA. IEEE.

Henry Moss, David Leslie, and Paul Rayson. 2018. Using J-K-fold cross validation to reduce variance when tuning NLP models. In *27th International Conference on Computational Linguistics*, pages 2978–2989, Santa Fe, New Mexico, USA. ACL.

Justyna Olko and John Sullivan. 2016. Bridging gaps and empowering speakers: An inclusive, partnership-based approach to Nahuatl research and revitalization. *Integral strategies for language revitalization*, pages 347–386.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *2014 EMNLP*, pages 1532–1543. ACL.

Robert Pugh, Cheyenne Wing, María Ximena Juárez Huerta, Ángeles Márquez Hernandez, and Francis Tyers. 2025. Ihquin tlahtouah in tetelahtzincocah: An annotated, multi-purpose audio and text corpus of western sierra Puebla Nahuatl. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3549–3562, Albuquerque, New Mexico. ACL.

Mitsuya Sasaki. 2022. Divide y entenderás: El papel de la polarización sintáctica en el náhuatl moderno y colonial. In *Coloquio de Investigación Lingüística, Universidad de Sonora (Mexico)*.

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

Juan-Manuel Torres-Moreno, Juan-José Guzmán-Landa, Graham Ranger, Martha Lorena Garrido-Avendaño, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Carlos-Emiliano González-Gallardo, Elvys Linhares Pontes, Patricia Velázquez Morales, and Luis-Gil Jiménez-Moreno. 2024. π-yalli: un nouveau corpus pour le nahuatl. *arXiv:2412.15821*.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media.

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