μGNAW$\oplus$(0,1,2): three context-free micro-grammars oriented for machine learning of Nawatl language

Anonymous ACL submission

Abstract

We present in this article three context-free micro-grammars (CFMG) for the Nawatl language. Nawatl, an Amerindian language, is a π-language, or a low-resource language in terms of digital data, and the corpora available for Large Language Models (LLMs) are virtually nonexistent. The objective is to generate a significant number of Nawatl sentences in order to expand the available corpus for training static embeddings or LLMs. Using the best micro-grammar, we significantly expanded the Nawatl corpus π-YALLI, then we used this enriched corpus to train FastText and applied it to a sentence-level semantic task. The results show an improvement compared to results obtained using only the original corpus without artificial expansion, which is encouraging.

1 Introduction

Nawatl (or Nahuatl) is one of the indigenous languages of Mexico and the most widely spoken national indigenous language, 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 Huastec 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 significant variation in orthography and lexical choice (Zimmermann, 2019),(Olko and Sullivan, 2016),(Hansen, 2024).

The goal of the proposed grammars is not to model fully the complexity of the Nawatl language, 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 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 the Nawatl corpus π-YALLI (Guzmán-Landa et al., 2025), which has previously been employed in word-level semantic similarity tasks. The vocabulary size has a direct impact during training: a larger vocabulary reduces the average sequence length but increases the size of the embedding matrix, which is expected to positively affect model performance (Tunstall et al., 2022), (Goyal et al., 2018).

The structure of the paper is as follows: Section 2 provides an overview of the Nawatl language and its grammatical features. Section 3 introduces context-free grammars. Section 4 presents the proposed context-free micro-grammar for Nawatl. Section 5 describes the extended corpus, π-yall-IA. Section 6 reports on experiments using the micro-grammars in a semantic similarity task. Finally, Section 7 concludes the paper and outlines directions for future work.

2 Nawatl language

Nawatl, as 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 this allows 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

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

well as information about the actants, modal, directional, and relational elements (see (Launey, 1978), (Flores Nájera, 2019), (Sasaki, 2022)).

	Example
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: Examples of syntactic constructions in Nawatl. More frequent structures are indicated in bold.

3 Context-Free Grammars (CFGs)

A context-free grammar is a type of 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 or more) formed with symbols from V and Σ; and $S \in V$ 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.

4 Two Context-Free Micro-grammars for Nawatl: μGNAW$\oplus$(0,1)

In this section, we introduce two context-free micro-grammars for Nawatl. These grammars are considered micro versions primarily because both avoid the use of recursive production rules. Similarly, only the first three grammatical persons (I, you, he/she, it) are included, verbs being limited to the singular form and present tense.

4.1 A first context-free micro-grammar for Nawatl: μGNAW$\oplus$0

Our initial approach to Nawatl grammar (see Table 2) μGNAW$\oplus$0, is inspired by the grammatical frameworks developed for Indo-European languages. Given that Nawatl belongs to the families of indigenous American languages —linguistically distant from Indo-European families— 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: $P \rightarrow$ noun phrase (N) verb phrase (V), where V may also contain N is insufficient to capture the deeper syntactic organization of Nawatl, which is centered around the verb. We therefore focus on a classical model consisting of two primary phrase types: the noun phrase (N) and the verb phrase (V). These phrases may include additional elements (particles, nouns, and prefixes) that are translated into our grammatical categories as temporal and quantifying adverbs (ADV_T, ADV_Q), adjective nouns (ADJ), personal (PP) and person markers (PV), possessive markers (POS), and negation (NEG). Although derived from the grammars of families of τ-languages, this preliminary version nonetheless permits the generation of some basic Nawatl structures.

For the sake of simplifying our approach, in the two micro-grammars we will use only the singular grammatical persons. The plural will be left for a later work.

```
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 travel | to eat |
to listen
⊕ = concatenation ∅ = null _ = space
```

Table 2: Grammar μGNAW$\oplus$0.

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 your rabbit*

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

- $PP_{1,2,3}$ $PV_{1,2,3} \oplus v$: na/ta/ya ni/ti/_/nehnemi *# I am/you are/he/she is (who) I/you/he/she walks*
- $POS_1 \oplus n$ $PV_3 \oplus v$: nomiston _nehnemi *# my_cat it_walks*
- $POS_2 \oplus n$ NEG $PV_3 \oplus v$: momiston amo _nehnemi *# my_cat it doesn't walk*
- PP_1 NEG $PV_1 \oplus v$ ADV_Q n: na amo nikkwa miyak xokotl *# It's me (who) I don't eat a lot [of] fruit*
- ADV_T $ADJ(ART\ |POS) \oplus n$ NEG $PV_3 \oplus v$ 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*

4.2 A more realistic context-free micro-grammar for Nawatl: μGNAW$\oplus$1

Our second version of the Nawatl grammar is grounded in the prototypical grammatical structures characteristic of VSO (and derived) word order, as well as the less frequent SVO patterns outlined in Section 2.

The primary aim of this revised grammar is to accurately model the various types of verb phrases (V), which are frequently employed, alongside the comparatively less common noun phrases (S). These phrases may incorporate adjunct elements such as adverbs, person markers (PV), object markers (PO), possessive markers (POS), and negation (NEG), in addition to terminal nodes representing nouns (n), verbs (v), and an auxiliary verb. To achieve a closer approximation to authentic Nawatl grammar, the following modifications have been implemented:

- The number **se** (one) and the particle **ni / in** (the) have been omitted;
- The person markers **na / neh** (I/me), **ta / teh** (you), and **ya / yeh** (he/she/it) have been excluded;
- Adjectives have been reanalyzed as nouns;
- Quantifying and temporal adverbs have been reclassified as intensity adverbs (ADV_I);
- Locative adverbs (ADV_L) have been introduced;

Thus, this micro-grammar has been formulated to reflect the structural patterns observed in Nawatl as employed by native speakers. Within the grammar μGNAW$\oplus$1 (see Table 3), the verb assumes a central and predominant role in the syntactic constructions of Nawatl.

```
P  →  VSO|VO⃗  |  ⇠VS|VOS|SV|SOV|SVO
V  →NEG ADV_I  PV_i⊕PO_j⊕v ; i , j = 1 ,2 ,3 ; i≠j
S  →  POS⊕n  ADV_L
O  →  S
POS  →  no−|mo−|i −|∅  # my|your|his ,her , its
NEG  →  amo|∅    # no/not
ADV_I  →  tlawel|miyak|nochi|∅
          # too much|a lot|all|∅
ADV_L  →  −kan|−ko|−ihtik|∅
          # place|inside|the inner part|∅
PV_i  →  ni −|ti −|∅− # I,you,he/she/it
PO_j  →  nech −|mits −|ki −|tla −
         # to me|to you|to her/to him|to it
v  →  nehnemi|kwa|kaki # to travel|to eat|
      to listen
n  →  siwatl|miston|elotl|xokotl|tochin
      |yolkatl|nakatl # woman|cat|corn|
      fruit|rabbit|animal|meat
⊕ = concatenation ∅ = null
```

Table 3: Grammar μGNAW$\oplus$1.

The $\overrightarrow{VO}$ and $\overleftarrow{VS}$ respectively indicate reasonably the reading direction of the structure —either right to left or left to right— to avoid or at least, contextto minimize, the ambiguities between the subject and the object. For example, the VS sentence:

$\overleftarrow{\textit{Kitta tlakatl}}$: A man sees (him, her, it)

in opposition to the VO sentence:

$\overrightarrow{\textit{Kitta kalli}}$: (he/she) sees a house

Rhetorical connectors CR (*and, but, moreover, however*, etc.) would add greater variability and richness to the generated sentences, and could be incorporated into the presented grammar as follows: $P \rightarrow P\ CR\ P$. However, due to its recursive nature, this rule will not be included in the current version.

Some examples illustrating the various structures of the grammar μGNAW$\oplus$1 are presented in the Table 1. Of course, to remain realistic, in both grammars μGNAW$\oplus$(0,1), the lists of nouns N and verbs V must include a sufficiently large number of distinct lexical entries.

4.3 Computing the number of sentences

We have estimated the number of sentences $\mathfrak{F}^0$ generated by the grammar μGNAW$\oplus$0. This number was calculated by considering the combinatorial possibilities of the grammar's rules and elements, as follows:

$$\mathfrak{F}^0 = ADV_T\,\{ADJ \cdot N \cdot (ART + POS) \cdot [1 + NEG \cdot V \cdot ADV_Q] + PP_i \cdot V \cdot NEG \cdot ADV_Q \quad (1)$$

Similarly, for the grammar μGNAW⊕1, the expression for the calculation of the number $\mathfrak{F}^1$ is:

$$\mathfrak{F}^1 = (NEG \cdot ADV_I \cdot v)\,\{4(POS \cdot N \cdot ADV_L)\,[NEG \cdot ADV_I \cdot v + POS \cdot N \cdot ADV_L] + 2POS \cdot N \cdot ADV_L + NEG \cdot ADV_I \cdot v + POS \cdot N \cdot ADV_L \quad (2)$$

The following calculations show the values estimated of $\mathfrak{F}^0$ and $\mathfrak{F}^1$ in relationship with the terminal symbols coming from the grammars (Table 2 and Table 3) or from the Aulex[2] dictionary:

μGNAW⊕0 : N=7, V=3, ADV$_Q$=3, POS=3, ART=3, ADV$_T$=4, ADJ=3, PP=3, NEG=3: $\mathfrak{F}^0$=14,436.

Aulex : N=11178, V=2223, ADV$_Q$=417, POS=3, ART=3, ADV$_T$=4, ADJ=3, PP=3, NEG=2: $\mathfrak{F}^1$=1.49×10^{12}.

μGNAW⊕1 : n=7, v=3, PO$_j$=4, PV$_i$=3, ADV$_L$=4, ADV$_I$=3, POS=4, NEG=2: $\mathfrak{F}^1$=14 547 456.

Aulex : n=11178, v=2223, PO$_j$=4, PV$_i$=3, ADV$_L$=40, ADV$_I$=76, POS=4, NEG=2: $\mathfrak{F}^0 = 2.63 \times 10^{20}$.

Finally, as a scientific curiosity, with the grammar μGNAW⊕1, using the number of N, V, and $ADV_{L,I}$ from the Aulex dictionary, the number of productions $\mathfrak{F}^1 \approx 2.63 \times 10^{20}$ (far fewer than a googol).[3] However, if a *CR* rhetorical connector is inserted between two phrases, the potential number of possible phrases is modified as follows $\mathfrak{F}^1 \times \mathfrak{F}^1 = \mathfrak{F}^{1^2} \approx (2.63 \times 10^{20})^2 = 6.2 \times 10^{40}$. Five rhetorical connectors enable one to reach orders of magnitude beyond a googol $\mathfrak{F}^{1^5} \approx 1.2 \times 10^{102}$. Now, considering the combinatorial potential of the 25 rhetorical connectors available in the Aulex dictionary, the number of productions is $(2.63 \times 10^{20})^{25} = 3,15 \times 10^{510} \approx 5$ googol.

We chose to implement both grammars μGNAW⊕(0,1) in Prolog, language which is very expressive and well-suited to the generation of context-free grammars. Moreover, Prolog is fast and handles memory efficiently. Indeed, the generation of more than 14.5 million sentences of μGNAW⊕1 required between 30-50 seconds on a GNU/Linux machine with an i5 processor and 16Gb of RAM.

The production of grammatical sentences requires a minimal post-processing step, which involves: token concatenation, removal of duplicate letters, normalization of spaces, capitalization of the first letter, and insertion of a final period.

In this way, we have implemented a sequence-to-sequence model to produce a syntax structure V, VS, VO, VSO, and SV as predicative, simple, juxtaposed, or subordinate sentences part of the grammar μGNAW⊕1, but after we focus on the structure VSO and even VOS, reducing complexity and gaining grammaticality in the generated sentences, the better to understand and assess the impact on the π-YALLI corpus.

5 π-YALL-IA, an Artificial Nawatl corpus

We observed that the two grammars μGNAW⊕(0,1) are capable of generating a large number $\mathfrak{F}^{0,1}$ of grammatically acceptable productions. However, although the values of $\mathfrak{F}^{0,1}$ are large, they represent only a tiny fraction of the number of sentences that could be produced using recursive grammars. Such recursive grammars, however, fall outside the scope of this article. Nevertheless, the number of sentences $\mathfrak{F}^{0,1}$ 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 walks too much*" because, although they belong to μGNAW⊕(0,1), they are not semantically realistic. We must therefore implement filters.

5.1 Semantic Filters

The artificial corpus derived from the grammar 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:

- A filter that attempts to approximate semantics using supervised classification methods.

[2] https://aulex.org/nawatl/

[3] A googol = 1×10^{100}, https://fr.wikipedia.org/wiki/Gogol_(nombre)

- A filter based on the association between verbs and animate/neutral/inanimate nouns.
- No nouns repeated in the same sentence.
- In the case of μGNAW$\oplus$1, two more filters concerning the VSO structure, were applied.

The semantic filter based on supervised learning is computationally greedy and will be implemented in future work. The animate/neutral/inanimate filter enables the elimination of a significant subset of sentences that lack reasonable semantics at low computational cost. Some examples of the filter can be seen in appendix 1. Furthermore, given the nature of grammar 1, it allows us to apply two extra filters, for example, when we have VSO where S and O cannot be identical.

This approach enabled us to augment the π-YALLI corpus with an additional number of N tokens, while still respecting a plausible Nawatl grammar.

From the grammars μGNAW$\oplus$(0,1), we generated sets of $\mathfrak{F}^{0,1}$ sentences. These sets, once appropriately filtered, yielded two versions of the π-YALLIA$\oplus$(0,1)* corpus, containing respectively $\mathfrak{F}^{0,1}*$ sentences. The semantic approximation filter associates nouns with an animate or living referent (e.g. animals, people) to verbs expressing appropriate actions (e.g. to walk, to eat, to fly), in contrast to inanimate referents (e.g. things, artifacts, phenomena) and their corresponding verbs.

This procedure allowed us to retain only a set of $\mathfrak{F}^{0,1}*$ grammatically correct and semantically acceptable sentences, resulting in an artificial and usable corpus of Nawatl sentences. In appendix 2 we present several examples for our grammars.

Accordingly, μGNAW$\oplus$(0,1), we obtained:

- $\mathfrak{F}^0* = 6$K sentences with both filters applied ($14,664$ without filtering);
- $\mathfrak{F}^1* = 1$M sentences with all four filters applied ($15,482,880$ without filtering).

The values used in the equations (1) and (2) to obtain the number of sentences generated by both grammars are presented in the Tables (4) and (5).

N	V	ADV$_Q$	POS	ART	ADV_T	ADJ	PP	NEG	$\mathfrak{F}^0$
8	7	5	3	3	8	1	3	1	14.7K

Table 4: Knowledge base for μGNAW$\oplus$0.

N	v	PO$_J$	PV$_I$	ADV$_L$	ADV$_I$	POS	NEG	$\mathfrak{F}^1$ VSO
8	7	4	3	6	5	4	1	15.5M

Table 5: Knowledge base for μGNAW$\oplus$1.

5.2 Merging artificial sentences and the Nawatl π-YALLI corpus

The idea that emerges now is the union of an "authentic" corpus such as π-YALLI with an artificially generated corpus. We think that this union may be beneficial for teaching Language Models (whether static or LLM). In particular, we have focused on static LM training. After the union of both corpora, the new corpus goes through an elementary process of unification of spellings, in order to standardize the use of characters of the Nawatl variants.[4]

We now present some characteristics of the –authentic– π-YALLI Nawatl corpus (Guzmán-Landa et al., 2025), that we have used for our semantic similarity experiments. The corpus is heterogeneous in terms of categories and linguistic variants in Nawatl, and it contains a relatively small number of tokens ($\approx$ 6.1M) and sentences ($\approx$ 300K). This makes it useful for training static LLMs, but clearly inadequate for training dynamic LLMs. Indeed, it has been reported that dynamic LLMs require in the order of 10M-100M tokens to achieve stable embeddings (Micheli et al., 2020). For this reason, we decided to unify the π-YALLI corpus with the filtered artificial sentences generated by the grammars. The new corpora, π-YALL-IA$\oplus$(●), were used to train FastText static embeddings with the objective of improving performance on the task of sentences semantic similarity.

6 Experiments: a sentence-level semantic similarity task

In this section, we present a protocol to evaluate our approach. Semantic similarity is a classical 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 gives rise to a ranking of candidates that can be compared against a human-generated ranking. This is the

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

evaluation protocol used in (Guzmán-Landa et al., 2025), which we adopt here by training embeddings to assess the impact of learning based on both existing and artificially generated corpus.

Figure 1 illustrates the complete experimental pipeline we designed, from the construction of the corpus (originating from authentic documents or formal grammars) to the evaluation of the semantic task via a reference ranking produced by human annotators, compared against the ranking generated by FastText.

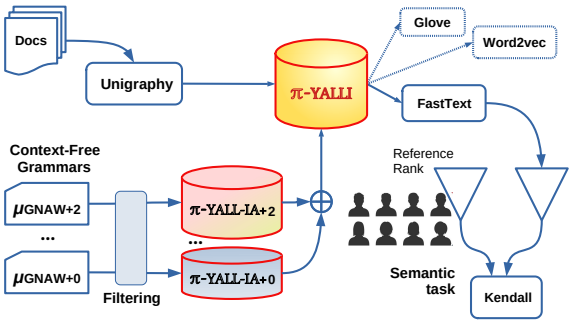

Figure 1: Outline of the semantic similarity task.

We did not expect to achieve high performance on the semantic task using only the artificial corpus generated by the μGNAW$\oplus(\bullet)$ grammars. Rather, our goal was to enrich the π-YALLI corpus with these new artificial datasets. Our hypothesis is that the μGNAW$\oplus$1 corpus—which adheres more closely to Nawatl grammar—should contribute to improving the performance of static language models.

The extended corpora π-yall-IA$\oplus(\bullet)$ were thus used to train FastText static language models (Bojanowski et al., 2017) from scratch. The resulting embeddings were then evaluated on a semantic sentences similarity task. The sentence semantic similarity task used is the following: given a set of reference sentences r_i, $i = 1, 2, ..., R$ (without context), where the reference of each sentence is associated with a list L of 5 candidate sentences. The goal is to sort the list L according to the semantic similarity to its corresponding reference sentence. A total of $R = 30$ references were written in Spanish by an annotator of our team. 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 semantic similarity to the original sentence. All sentences (references and candidates) were translated into the Central Nawatl variant by a bilingual Nawatl speaker (Spanish-Nawatl) of our team. The basic statistics for this task are as follows:

	Sentences	Tokens	Types	Mean
Refs	30	246	189	8.10
Cands	150	1026	599	6.84

Appendix shows two examples of reference–candidate sentence blocks that are included in this reference ranking (R_R). 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 similarity measure between rankings is the Kendall's $\tau(R_R, R_M)$ (Kendall and Smith, 1939)[5].

6.1 The LLMs employed

In this work we have tested the following five LLMs models via APIs[6]: ChatGPT-4 mini API; Gemini-2.5-flash-preview-05-20 API[7]; DeepSeek-V3-0324 API[8]; Llama-3.1-70B-Instruct API[9] and Mistral-large-latest API[10]. In an interactive user mode we have employed three models: Copilot (`https://www.microsoft365.com`), Grok 3 (`https://grok.com`) and Claude 3.7 (`https://claude.ai`). The prompt used was the following:[11] *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.* The references and candidates were written in Central Nawatl.

6.2 Results and Discussion

Table 6 summarizes all our results. As shown in this Table, the best result was obtained using the filtered π-YALL-IA$\oplus$1 corpus, which yielded a Kendall's τ of

[5]Available at `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html`

[6]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.

[7]`https://deepmind.google/models/gemini/flash/`

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

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

[10]`https://mistral.ai/news/mistral-large`

[11]The prompt was written in French: « *À partir de la phrase nahuatl: "[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.* »

0.506. This places the FastText model trained with an expanded corpus in third place in the ranking, only behind the LLMs Gemini 2.5 and Claude 3.7.

Algorithm	τ
Gemini 2.5-flash-preview-05-20 API	**0.653**
Claude 3.7	0.566
FastText Skip π-YALLI $\cup$ μGNAW$\oplus$1, 4 SW	***0.506***
DeepSeek V3-0324 API	0.500
FastText Skip π-YALLI $\cup$ μGNAW$\oplus$2, 4 SW	***0.493***
Grok 3	0.486
FastText Skip π-YALLI, 4 SW	***0.480***
FastText Skip π-YALLI $\cup$ μGNAW$\oplus$0, 4 SW	***0.473***
Copilot	0.446
Mistral-large-latest API	0.360
ChatGPT-4 mini API	0.327
Llama-3.1-70B-Instruct API	0.219

Table 6: Semantic similarity task: LLMs vs FastText. Stopwords eliminated: *iwan, in, tlen*, and *ipan*.

As a peculiar result, using μGNAW$\oplus$1 grammar (that avoid the typology of transitive and intransitive verbs and the impact of the possessive on subjects), FastText achieves a performance of 0.506. However, the sentences generated in this way are mostly ungrammatical and meaningless. We then decided to refine the grammar μGNAW$\oplus$1 to decrease the number of ungrammatical sentences produced and to establish syntactical relationships between words in its logical composition. To this end, we considered six important elements:

- Exclusive implementation of base 1 verbs.[12]
- Exclusive use of transitive verbs and verb conjugation and valency in the 3rd-person singular.
- Integration of a larger number of nouns, especially animate nouns because they associate well with transitive verbs.
- Intensity adverbs are excluded.
- Only the OR structure can have an adverb of location after a noun n.
- Adjectives are part of the S and O structures, i.e. they are positioned before the possessive POS.

[12]Verbs in Nawatl can have 3 bases, base 1 represents all those verbs in the present-imperfective tenses, base 2 the verbs in the perfective tenses and finally, base 3 represents the verbs in the future-hypothetical tenses.

Using this new grammar (see Table 7), which we will call μGNAW$\oplus$2, we manage to obtain grammatically correct sentences and a Kendall's $\tau = 0.493$.

```
P   →  VSO | VO⃗ | V⃖S | VOS | SV | SOV | SVO
V   →  NEG PV_i⊕PO_j⊕v ;  i,j=1,2,3; i ≠ j
S   →  ADJ POS⊕n
O   →  ADJ POS⊕n ADV_L

ADJ   →  weyi | istak | ∅     # big | white
POS   →  no- | mo- | i- | ∅  # my | your | his, her, its
NEG   →  amo | ∅              # no/not
ADV_L →  nikan | nepa | ∅     # up | there | ∅
PV_i  →  ∅-                   # he/she
PO_j  →  ki-              # to her/to him
v  →  # List of transitive verbs
n  →  # List of nouns
⊕ = concatenation  ∅ = null
```

Table 7: Grammar μGNAW$\oplus$2.

Computing the ratios $\rho_{(0,1,2)}$ (the proportion of artificial tokens produced by the grammars) with respect to the π-YALLI text, we obtain:

- ρ_0 (π-yall-IA$\oplus$0) = $\frac{25K}{6.1M} \approx 00.40\%$
- ρ_1 (π-yall-IA$\oplus$1) = $\frac{6.17M}{6.1M} \approx 100.92\%$
- ρ_2 (π-yall-IA$\oplus$2) = $\frac{5.42M}{6.1M} \approx 88.67\%$

We thus observe that the ratio ρ_0 corresponding μGNAW$\oplus$0 represents too low proportion of artificial text, by introducing negligible grammatical and semantic information into the model. This may explain the low performance of the FastText model ($\tau = 0.473$), which is even lower than that obtained without the addition of artificial text ($\tau = 0.480$). This phenomenon does not occur in the case of μGNAW$\oplus$(1,2), where the authentic/artificial tokens ratios $\rho_{1,2}$ are more important and comparatively equivalent. Indeed, our hypothesis regarding the expansion of the Nawatl corpus proved more reliable with grammar μGNAW$\oplus$2. This suggests that a more accurate modeling of Nawatl grammar does help to capture the deeper structures of the language— such as agglutination, polysynthesis, and the central role of the verb— resulting in a (slight, yet directionally meaningful) improvement in performance on the semantic similarity task involved.

7 Conclusions and Future works

The generative use of the proposed μGNAW$\oplus$(0,1,2) grammars allows the creation of a large number $\mathfrak{F}^{0,1,2}$ of sentences, that may be considered as –very

large– artificial corpus. A semantic filtering process was applied to these artificial corpus in order to reduce the universe to $\mathfrak{F}*^{0,1,2}$ sentences. Subsequently, the artificial corpus were concatenated with the authentic π-YALLI corpus. The corpus enlarged resulting from the union of μGNAW$\oplus$2 (that is more closer to the Nawatl language) and π-YALLI, named as π-YALL-IA, containing over 10.79M tokens ($\approx$ 1.28M sentences), allowed for a more efficient training process of static embedding models. Using π-YALL-IA, FastText was able to improve their performance on the semantic sentence similarity evaluation task, raising τ from 0.480 to 0.493.

The approach to Nawatl grammar using a CFG micro-grammar in generative mode, seems to help the learning of FastText-like static models. The resulting embeddings better capture the structure of the Nawatl language. However, there are still difficulties and challenges in relationship to the deep structure of an agglutinative language, morphophonological modifications of roots by composition, and features of dialect variation. Even though different syntactic structures can be produced and thus new grammatical sentences generated, the complexity of syntactic relationships must be explored in depth. These difficulties are lessened in the case of processing isolating or moderately synthetic languages such as Indo-European languages. Thus, the Nawatl as polysynthetic languages is a complex morphological case that requires, even for humans(Kelly et al., 2014), the ability to identify roots on which to incorporate multiple affixes, but also to confront the rules of composition and, above all, phonetic and morphological modifications. To all this, we must add the dialectal diversity of the forms. Thus, in our case the problem is to achieve a morphological regularity and not a morphological complexity as the relevant factor for learning, since this kind of language may be relatively regular in the templatic sequence used (Kelly et al., 2014).

In future work, we will seek to increase the number of n and v elements, include additional grammatical persons, verb tenses, plurals, rhetorical connectors, and improve the semantic filter (to avoid effects of combinatorial explosion). Another idea to try is the introduction of Probabilistic Context-Free Grammar (PCFG), where nodes can have associated probabilities. This kind of grammars[13] (Smith and Johnson, 2007) is particularly useful when filtering semantic/non-semantic phrases using an external classifier. This classifier starts with manual labeling to produce automatic labeling of a subset of phrases. Such classification could be used to modify the parameters (in this case, the probabilities) of the generative grammar. On the other hand, this effort adds to the researches that, despite being highly complex and low-resource languages, recognize the broad possibilities for computational development with the languages of the Uto-Aztecan family (Mager Hois, 2024), (Eskander et al., 2019). Thus, it is also important to introduce an extended recursive grammar capable of modeling the structures produced by native Nawatl speakers more realistically. We also intend to evaluate the contribution of the artificial corpus to a Sentiment Analysis task, Text Summarization (Torres-Moreno, 2014) and Named Entities recognition.

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

8 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.

9 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.

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.

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 ACL*, 5:135–146.

Ramy Eskander, Judith L Klavans, and Smaranda Muresan. 2019. Unsupervised morphological segmenta-

tion for low-resource polysynthetic languages. In *16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195.

Lucero Flores Nájera. 2019. La gramática de la cláusula simple en el náhuatl de tlaxcala. *Recurso impreso, recurso electrónico*.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *2016 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.

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. 2025. π-yalli: un nouveau corpus pour le nahuatl / *Yankuik nawatlahtolkorpus pampa tlahtolmachiotl*. In *TALN 2025, Marseille*.

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

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`.

Barbara Kelly, Gillian Wigglesworth, Rachel Nordlinger, and Joseph Blythe. 2014. The acquisition of polysynthetic languages. *Language and Linguistics Compass*, 8(2):51–64.

M. G. Kendall and B. Babington Smith. 1939. The problem of m rankings. (3):275–287.

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.

Jesus Manuel Mager Hois. 2024. *Low resource NLP for polysynthetic languages: morphological segmentation and machine translation*. Ph.D. thesis, Dissertation, Stuttgart, Universität Stuttgart, 2024.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. *Preprint*, arXiv:2010.03813.

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.

Mitsuya Sasaki. 2022. Divide y entenderás: El papel de la polarización sintáctica en el náhuatl moderno y colonial. *Recurso impreso*.

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. 2014. *Automatic Text Summarization*. Wiley, London.

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.

Appendix 1: Animate/inanimate Filtering

This filtering works by tagging nouns and verbs with one of the tags already described. To illustrate how this filter works, we presents some examples of suitable verbs and nouns. 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.

Nouns: nantzin = mother (**animate**); mihkailwitl = day of the dead (**inanimate**); tatzin = father (**animate**); mapachin = raccoon (**animate**); kuawitl = eagle (**animate**).

Verbs: miki = to die (**animate**); ixpoliwi = to disappear (**animate/inanimate**); pia = to have (**animate/inanimate**); itta = to see (**animate**); chihua = to make (**animate**).

The rule says that if the labels match the sentence can be generated, but if the label is neutral, then a sentence with a animate and inanimate can be generated. That is, father (**tatzin**) and mother (**nantzin**) can die (**miki**), but not on the "Day of the Dead" (**mihkailwitl**), however, everyone can disappear (**ixpoliwi**).

Appendix 2: Examples of artificial sentences generated

This appendix shows some examples of artificial sentences generated by the grammars μGNAW$\oplus$(0,1,2). Starting with μGNAW$\oplus$0, which generates 6K sentences after passing through all the necessary filters. Here are some examples:

- *Aman ni kokostli miki miyak.* | Today the desease dies a lot.
- *Aman ni kokostli kahsikamati miyak.* | Now the desease understands a lot.
- *Axkan monakas xikkakikan nochi.* | Now it is your ear, listen to everything.
- *Niman isiwatl kahsikamati achi.* | Later, his woman understands enough.
- *Tonalli ta timotta achi.* | It is a day, you are, you see yourself enough.

In the case of the grammar μGNAW$\oplus$1, where we have 1M generated sentences, most of them agrammatical. Here some examples:

- *Miyak nechmiki nokokostli ahko mosiwatl tlakpak.* | A lot, to me, he/she dies, it is a bird/ plant/ desease that is on top, your woman is up.
- *Nochi mitsxikkakikan nonakas tlakpak motiotzin ompa.* | All, to you, listen my ear on top, it is your God there.
- *Achi kionikkak nokokostli oncan monantzin ahko.* | It is suficient, to him/her/it, I heard my bird/plant/desease there, you mother is on top.
- *Miyak tlakahsikamati ikokostli ompa tatzin nepa.* | His/her bird/ plant/ desease understands very much there a father is further.
- *Seki nechxikkakikan nonakas oncan itiotzin nepa.* | Someone to me. Listen my ear there his/her God is further.

Finally, the grammar μGNAW$\oplus$2, generates 957K artificial sentences, by example:

- *Kitoka istak motlaka weyi imapach nepa.* | Your white man follows his big raccoon there.
- *Kiitta miyak sentilistli weyi moxochih.* | Lots of families see your big flower.
- *Kimaka miyak imapach weyi notelpoch nikan.* | Here my big boy gives many of his raccoons.
- *Kiwelitta miyak moyolka weyi notelpoch nepa.* | Many of your farm animals like my boy there.
- *Kipaktia miyak isentilis istak moposol.* | Many in your family like white pozole.