# The Power of Simplicity: N-Grams and Transformers in Nahuatl Language Identification

**Luis Armando Mercado-Campos**[1] and **Robert Pugh**[2] and **Alexis Palmer**[1]

[1]University of Colorado Boulder, Department of Linguistics

[2]Indiana University Bloomington, Department of Linguistics

{lmercadocampos, alexis.palmer}@colorado.edu, pughrob@iu.edu

## Abstract

In the context of real-world language technology applications, the language or variety in which a given text is written is often unknown or uncertain. Yet, this information is crucial in order to adequately select and apply appropriate models or resources. Language identification (LID), or the process of determining the language or variety of a text sample, is thus often an important fundamental task in natural language processing. LID can be particularly challenging when: (1) there are not many labeled texts for training; and (2) similar or related languages are involved, since these may share a number of surface-level features. In this paper, we present an LID system for Nahuatl, a group of closely-related language varieties spoken in Mexico and Central America. Nahuatl LID involves both of the aforementioned challenges: Nahuatl varieties can be quite similar, sharing morphemes and even many lexical items, and there is a relative paucity of representative, variant-labeled Nahuatl text. We describe LID experiments for a total of 11 Nahuatl varieties, achieving generally good results (90.59% $\pm$0.09% in 5-fold cross-validation experiments). Many of the outstanding errors are the result of confusion between three highly-similar Huasteca variants.

## 1 Introduction

Nahuatl is a group of endangered, closely-related, morphologically rich languages[1] in the Nahuan branch of the Uto-Aztecan family, spoken primarily in Mexico. There are 30 formally recognized variants of Nahuatl (INALI, 2009), which can have substantial phonological, morphological, syntactic, and lexical differences (Gruda et al., 2023). There is also significant variant-internal diversity.

Figure 1: Map showing the geographic distribution of Nahuatl varieties in Mexico. Green, blue, and orange regions correspond to Central, Eastern, and Western varieties, respectively. Adapted from Pugh and Tyers (2024a).

Research on Nahuatl dialectology dates back to at least Lehmann (1920). Since then, researchers generally agree with the variant sub-classifications presented in Lastra (1986), Canger (1988), and Kaufman (2001). While not identical, these three agree on the existence of Eastern varieties, corresponding to one wave of early migration; Central varieties, corresponding to the Nahuatl spoken in the central valley of Mexico; and Western varieties, including Nayarit/Durango Nahuatl. Pharao Hansen (2014) provides additional recommendations for the classification of Eastern and Central/Western varieties based on a survey of linguistic evidence.

Importantly, while Nahuatl languages share a non-trivial number of grammatical features and lexical items, attempts to homogenize or downplay the diversity of Nahuatl varieties can negatively impact revitalization efforts (Hansen, 2013), and also fail to acknowledge the real and often substantial mu-

[1]We use "language," "variant," and "variety" interchangeably, but avoid the use of the term "dialect." In Mexico, this distinction carries real social, educational, and political consequences, as historically the term "dialect" has been used as a tool to marginalize indigenous language speakers (Aguayo Rousell and Piña Osorio, 2016). In government documentation, schooling, and public discourse, referring to these languages as "dialects" can contribute to marginalization and under-resourcing (Parodi, 2011).



*Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 153–167
July 3, 2026 ©2026 Association for Computational Linguistics

tual unintelligibility between speakers of different regional varieties. For example, Huasteca Nahuatl and Isthmus Nahuatl speakers often cannot communicate effectively without prior exposure (de Suárez et al., 1986).

We note that this system is designed as a variety classifier; it assumes the input is already known to be Nahuatl, and should be understood as operating downstream of a general language identification stage. Integrating it with a broader LID system (e.g., one that first filters for Nahuatl before variety classification) is a natural direction for deployment.

In a linguistic context like that of Nahuatl, a reliable variety identifier can serve many purposes. It can help organize and search web-scraped text or mixed-variant corpora such as Gutierrez-Vasques et al. (2016a), facilitating access for speakers, educators, students, and researchers. In applied settings, variety identification can support corpus building and documentation workflows by speeding up metadata assignment and flagging potential mislabels for manual review. Furthermore, for applications like automated dialog systems, being able to determine a user's variant can help ensure that any generated content is aligned with the user's language variety.

### 1.1 Related work

LID for text is a well-established and widely-researched task in NLP (Jauhiainen et al., 2019). The Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) has heavily focused on distinguishing between similar languages and varieties, including it in numerous shared tasks over the years (Zampieri et al., 2014). Early approaches to this task involved leveraging lexical and character features with linear classifiers, often combined with multi-stage classification (e.g. first identifying the language, then classifying the variety) and ensemble methods (Zampieri et al., 2014, 2015). More recently, approaches typically rely on large, pretrained transformer language models like BERT (Devlin, 2018) or XLM (Lample and Conneau, 2019), though uses of, e.g. SVM or XGBoost classifiers, are still common (Aepli et al., 2023).

In the context of LID for Indigenous languages of the Americas, there has been less focus. Pugh et al. (2025a) explore the task focusing specifically on the languages of Mexico, finding that character n-gram features with Linear SVM provide a strong baseline.

Computational work on Nahuatl is limited but growing. Gutiérrez-Vasques (2018) explores automatic bilingual lexicon extraction, demonstrating alignment potential despite data sparsity. Other work on the language has explored computational approaches to orthographic variation (Gutierrez-Vasques et al., 2025; Guzman-Landa et al., 2025), morphological analysis (Maxwell, 2005; Pugh et al., 2021; Tona et al., 2023; Tyers and Pugh, 2023), syntactic analysis (Pugh et al., 2022; Pugh and Tyers, 2024b), and corpus creation (Gutierrez-Vasques et al., 2016a; Pugh et al., 2025b; Guzmán-Landa et al., 2025). With respect to computational Nahuatl dialectology, Farfan (2019) developed a finite-state morphological analyzer built from a grammar of Classical Nahuatl to explore points of convergence among contemporary written Nahuatl variants. Character language models have also been shown to be effective at measuring the similarity between Nahuatl variants and for language detection (Pugh and Tyers, 2021).

The most relevant work to the present paper is Guzmán-Landa et al. (2026), which investigates methods of Nahuatl variant detection on a custom corpus. Their corpus is not publicly available.

**A note on data availability:** Data sovereignty is an essential concern for many language communities, and especially many Indigenous communities where people speak endangered languages. The CARE framework (Carroll et al., 2020) encodes ethical data governance practices to support scientific advancement while both safeguarding and honoring Indigenous data sovereignty. The data we have collected includes a mix of publicly-available data and data collected and/or built through direct personal collaborations with speakers and speaker communities. We intend to release at least a portion of this corpus, and we are still in the process of discussing with our collaborators how best to do this, while respecting CARE and general principles of fair use.

### 1.2 Research question and contributions

The core of this work is the evaluation of models that learn variety-discriminating features directly from naturally occurring text, without relying on hand-crafted linguistic rules. Specifically, we investigate how effectively architectures like XLM-R and supervised machine learning models trained on character n-gram features (as well as ensembles of various models) can capture regional linguis-

tic patterns and orthographic variations to achieve high-accuracy language identification across a diverse linguistic continuum.

We introduce a system for automatic classification of Nahuatl variants, using a weighted soft-voting ensemble combining a linear SVM, a logistic regression model, and XLM-R. We show that this combination is effective for LID, achieving 91% accuracy on a held-out test set. Importantly, our model retains the orthographical variation seen in naturally-occurring data.

## 2 Linguistic properties of Nahuatl languages

Nahuatl languages are morphologically agglutinative, with extensive use of derivational and inflectional morphology. Canonical word order is typically SOV (Subject-Object-Verb), though constituent order can vary depending on pragmatic or regional factors (Olko et al., 2018; Gruda et al., 2023). Despite their shared history, Nahuatl varieties diverge significantly in phonology, lexicon, and syntax, which reflects both internal change and external influences from Spanish and neighboring Indigenous languages (Parodi, 2011).

Like other Uto-Aztecan languages, Nahuatl has a relatively small but distinct phonemic inventory. The overview in Table 1 represents a synthesized core inventory based on our analysis of both Classical Nahuatl (*nci*) as a stable reference system and documented modern regional innovations. While this work uses text rather than speech, phonological distinctions between varieties are directly relevant because they are often systematically reflected in orthography. For example, the 'tl-t-l' isogloss and the presence or absence of the saltillo (glottal stop) manifest as consistent spelling differences across varieties, making them useful discriminative features for test-based classification. This aggregated inventory draws specifically from the surveys of Canger (2001) and de Suárez et al. (1986) to establish the segments that participate in systematic regional mappings. Rather than reflecting a single 'standard', this inventory serves as a comparative work to ensure that phonological differences (e.g., 'saltillo' ʔ or the /t͡ɬ/ variants) remain available as discriminative features for the classification models.

| Place of Articulation | Stop | Affricate | Fricative | Nasal | Liquid | Glottal |
|---|---|---|---|---|---|---|
| Bilabial | p | - | - | m | - | - |
| Alveolar | t | ts, t͡ɬ | s | n | l, r | - |
| Palatal | - | t͡ʃ (ch) | ʃ (x) | - | - | - |
| Velar | k, k$^{w}$ | - | - | - | - | - |
| Glottal | - | - | h | - | - | ʔ |

Table 1: Aggregated consonant inventory across Nahuatl varieties, serving as a reference for understanding the phonological distinctions that may surface as discriminative orthographic features in text. Not all contrasts are present in every variety. This inventory represents the union of documented segments across varieties, with Classical Nahuatl (nci) as a reference point.

## 3 Data

We begin by assembling a corpus of over 1.1M sentence-level entries from eleven regionally distinct Nahuatl languages. he distribution of this data across varieties, sources, and domains is shown in Table 2, with additional details in Table 10 in the Appendix.

### 3.1 Corpus Sources and Description

Data is curated from a variety of sources, both publicly available and private. Only some of the private-source data will be made available, by agreement with those sources.

**(i) Existing Machine Translation Corpus.** Mercado-Campos (2023) compiled a corpus of approximately 31,000 regionally labeled sentences from openly-licensed datasets. This subcorpus includes texts from multiple genres such as dialogues, narratives, sermons, and health pamphlets. Although not structurally parallel, many of these documents address similar domains, particularly education and religious instruction.

**(ii) New Collaborator Contributions (2024-2025).** In ongoing collaborations, we have collected new data from several Nahuatl varieties. Most of this data has been shared informally through personal interactions, primarily via social media or local archives. This subcorpus includes a mix of narrative fragments, conversational transcripts, and educational material. Consent to use and share this data was granted by the language communities for this research project only.

**(iii) Personal Language Learning Collection.** As part of an ongoing language learning effort, new examples of one variety were collected from annotated lessons, beginner notes, and other class-related materials. Familiarity with the variety through active language study informed quality

| Lang | Name | # of Sentences | Domain |
|---|---|---|---|
| nhw | Western Huasteca Nahuatl | 369,830 | Daily Life, Bible |
| nhe | Eastern Huasteca Nahuatl | 168,799 | Legal, Stories, Bible |
| azz | Highland Puebla Nahuatl | 130,151 | Daily Life, Stories, Bible |
| ncj | Northern Puebla Nahuatl | 129,829 | Transcripts, Legal, Bible |
| nch | Huasteca Nahuatl | 105,856 | Educational, Myths, TikTok Stories, Bible |
| nhi | Western Puebla Sierra Nahuatl | 90,157 | Transcripts, Bible |
| nhg | Tetelcingo Nahuatl | 87,859 | Daily Life, Bible |
| nlv | Orizaba Nahuatl | 14,636 | Educational, Government, Bible |
| nhm | Morelos Nahuatl | 7,515 | Educational, History, Bible |
| nci | Classical Nahuatl | 5,466 | Poetry |
| nhn | Central Nahuatl | 3,062 | Stories, Daily Life, Children Stories |
| **Total** | | **1,113,160** | |

Table 2: Language variety statistics (ordered by # of sentences)

judgments during curation. The informal nature of this data required extra curation to be suitable for use in this project. The subcorpus consists of approximately 43,400 sentences, after filtering out low-quality, duplicate, or dictionary-style entries. We additionally exclude many examples extracted from images due to OCR errors or poor formatting.

**(iv) AmericasNLP Shared Task Data.** The AmericasNLP 2025[2] shared task competitions (De Gibert et al., 2025) include a small amount of non-parallel Nahuatl text (e.g., daily conversations) from 2 varieties.

**(v) Bible Corpus.** The massively parallel Bible corpus (Christodouloupoulos and Steedman, 2015) contains the New Testament (NT) for some Nahuatl languages. Between this source and a scripture website,[3] we collect NT Bibles for 16 different Nahuatl languages. There is a significant history of research using the Bible as a data source for very low-resource languages, to varying effect (among others, Chew et al., 2006; Mayer and Cysouw, 2014; Agić et al., 2015; Nicolai and Yarowsky, 2019; Ebrahimi and Kann, 2021; Liu et al., 2021; Kann, 2024; Marashian et al., 2025; Le Ferrand et al., 2025). Although the Bible is a readily-accessible source of parallel data for many languages, it must be used with caution. Researchers have raised concerns about the very particular domain of the Bible, the frequent use of archaic expressions, and – crucially – the fact that many early Bible translations were performed by colonial missionaries with only partial knowledge of the target languages.

**(vi) Axolotl Parallel Corpus.** Axolotl (Gutierrez-Vasques et al., 2016b; Gutiérrez-Vasques, 2018) is a freely-available parallel Spanish-Nahuatl corpus, accessible online and free to use either through it website or as a Python package (pyElotl Gutierrez-Vasques et al., 2025). Axolotl is also the Nahuatl name for the animal currently known as "ajolote" in Spanish or "axolotl" in English. Axolotl includes Nahuatl data from a range of variants and domains. Using pyElotl, we retrieve a subcorpus of about 13.5K sentences from 6 language varieties, summarized in Table 3.[4]

| Lang | # of Sentences |
|---|---|
| azz | 2,884 |
| nci | 5,421 |
| nhe | 149 |
| nhm | 1,938 |
| nhn | 1,757 |
| nhw | 1,449 |
| **Total** | **13,526** |

Table 3: Portion of Axolotl Corpus used for this study: # of sentences by variant

### 3.2 Variation within Variants: Orthographies, Registers, and Domains

Nahuatl materials span multiple orthographies and writing habits, adding to the variability seen even

[2]https://github.com/AmericasNLP/americasnlp2025/

[3]https://scriptureearth.org

[4]More detailed statistics in Table 11 in the Appendix.

within one variety. As an example, one important difference is the encoding of vowel length. Vowel length is phonemically contrastive in Nahuatl, meaning that short and long vowels distinguish meaning. Many of the varieties in our corpus use orthography to indicate vowel length, but there is no standardization across varieties in how specifically to write the contrast. Some orthographies mark long vowels with macrons (e.g., ā, ē, ī, ō, ū), others duplicate the vowels (e.g., aa, ee, ii, oo, uu), and others do not represent the contrast at all.

Our system is designed to perform variety classification without any orthographic normalization. We are, however, interested in understanding to what extent the model relies on orthographic differences for classification. To investigate that, we additionally produce a normalized version of the corpus using the `py-elotl` normalizer (Gutierrez-Vasques et al., 2025) set to the INALI standard, which maps orthographic variants to a single standardized form.

Additionally, our corpus includes a range of domains, as well as registers ranging from highly formal Bible translations and historical documents to contemporary daily-life conversations, social media data, and educational transcripts.

**Using Bible data.** To experimentally assess the effect of including Bible translations in datasets for this task, we create two additional subcorpora (see Section 5.2 for results and discussion). These datasets are restricted to the six languages for which we have both a Bible translation and non-Bible data, and each is balanced across languages. **Setting A** uses a balanced corpus of naturally occurring text, and **Setting B** adds a balanced selection of Bible data across the same languages.

### 3.3 Data Processing

**Unit of Analysis.** The goal in this project is language identification at the sentence level. We adopt a sentence-level approach for several reasons. First, many sources in our corpus do not have preserved document boundaries, making the document-level classification impractical. Second, some of the primary practical applications of this system (e.g., flagging mislabeled sentences in mixed-variety corpora) are inherently sentence-level tasks. Finally, in translated and parallel corpora, sentence-level correspondences across varieties are not guaranteed, as translators may convey the same meaning through different constructions entirely, making document-level context potentially misleading rather than helpful. Intuitively, we define "sentence" as a self-contained text unit conveying a single propositional idea, whether a standalone utterance or narrative fragment. This includes full constructions, short utterances common in dialogue, lines of poetry, and transcribed speech segments that may be grammatically incomplete but pragmatically whole. Some examples appear in Table 4.

| Category | Nahuatl | English |
|---|---|---|
| *Retained* | | |
| One-word sentence | *nimitstlasohtla* | I love you |
| Short utterance | *ninomachtia* | I study |
| Short utterance | *xinichpalewe* | Help me |
| Poetic line | *in xochitl, in kwikkatl* | the flower, the song |
| *Filtered Out* | | |
| Repetitive/Noise | *tla tla tla* | if if if |
| Unbound prefix | *nino-* | (reflexive prefix) I, me |

Table 4: Examples of sentence types retained and removed from the dataset.

We operationalize sentence segmentation utilizing line breaks, punctuation patterns, and formatting markers (e.g., chapter headings) as heuristics, followed by manual validation of the segmentation heuristics and spot-checking a random sample of sentences to verify data quality. Entries are cleaned of duplicate lines and dictionary-style lists, and highly fragmented tokens with insufficient context are removed to ensure data quality.

**Sampling and Balancing.** Distribution of sentences across language varieties is highly unbalanced (Table 2). While skew of this degree may reflect the prevalence of different varieties in accessible written media, it can obscure the true effectiveness of supervised classification systems. We thus create two versions of the corpus, one unbalanced and one balanced. Instances in the unbalanced version are split using stratified sampling to preserve the empirical regional proportions found in the raw corpus. To address potential domain and quantity bias, the balanced dataset is created by downsampling all varieties to match the smallest class (3062 sentences).

For experimentation, we reserve a held-out evaluation split using stratified shuffle sampling by language. Approximately 20% of the sentences for each Nahuatl variety are set aside as a test set, and the remaining 80% are used for training (and, where relevant, internal validation). This is applied

both to the unbalanced corpus and the downsampled balanced version, so that the label and domain distributions in the test set mirror those of the training data. All results reported are computed on these held-out test splits.

## 4 Models and Methodology

We train and compare several different classification models. We are specifically interested in combining the interpretability of non-neural models with the robust performance of pretrained multilingual models. To that end, we combine several models into an ensemble. All classification is performed at the sentence level.

### 4.1 Classification Models

We investigate three different models. First, we train a standard **logistic regression** model using TF-IDF weighted character n-gram features (n=2–5) to predict language variety given an input sentence. Second, using the same features, we train a **linear SVM** for the same task. Logistic regression is effective for sparse input features (Vimal and Anupama Kumar, 2020), while linear SVMs are particularly well-suited to capturing the orthographic and phonological cues highly predictive for Nahuatl (Çöltekin and Rama, 2016). Both model types have been shown to be strong, interpretable models for Indigenous language identification and classification. Together, these models serve as strong, interpretable baselines established for indigenous language identification and classification (Pugh et al., 2025a).

We compare XLM-RoBERTa Large (Conneau and Khandelwal, 2019) and mBERT (Devlin, 2018) as the multilingual baseline. Because of a mismatch between available sizes for these models, our comparison is between models of substantially different scales. XLM-RoBERTa Large (550M parameters), trained on 2.5TB of CommonCrawl data across 100 languages, is a more capable multilingual model for our task, and we find it outperforms mBERT-Large (330M parameters).

### 4.2 Model Ensemble

We combine the three models in a weighted soft-voting ensemble.

To determine relative weighting of models in the ensemble, we use a two-stage approach. First, we **optimize ensemble weights** using an 80/20 split of the training data. A grid search on this configuration determines the following weights: Linear SVM (0.60), Logistic Regression (0.20), and XLM-R (0.20).

Next, we perform a **robustness check** on the proposed ensemble weights using 5-fold cross-validation across the complete dataset. Models are retrained for each fold, using the ensemble weights above. Accuracies across the five folds range from 90.43 to 90.67. The stability of performance across folds verifies the stability of the proposed weights. NOTE: We performed the cross-validation experiments simply to validate our ensemble weights. No other results reported in this paper come from the cross-validation set up.

### 4.3 Tokenization

We compare SentencePiece (Kudo and Richardson, 2018) (Unigram LM (Kudo, 2018)), Byte-Pair Encoding (BPE (Sennrich and Haddow, 2015)), and character-level approaches. The custom SentencePiece Unigram tokenizer is selected based on classification perfomance in preliminary experiments (Macro F1 0.76, Accuracy 90%), as it best segments words into meaningful subwords while preserving orthographic patterns.

### 4.4 Implementation Details

Implemented using a single NVIDIA A100 GPU and Hugging Face Transformers. Maximum sequence length of 128. Training used AdamW optimizer, batch size 32, learning rate 2e-5, and early stopping based on validation loss with a patience of 3 epochs.

## 5 Results and Analysis

The results of the three individual models and the ensemble system on the held-out data are listed in Table 5. The system achieves an F1 score of 0.91.[5] It is worth highlighting that the ensemble system only slightly outperforms the much simpler Logistic Regression and Support Vector Machine models, a finding that supports the power of simple statistical modeling of character features for text-based LID.

The classification report shows strong performance for several more isolated languages (often approaching 0.99 recall), while the most challenging boundary is within the "Huasteca cluster" (*nch*, *nhw*, and *nhe*). In particular, *nhw* has the lowest

[5]This performance is consistent with the stable average found from the 5-fold cross-validation experiment.

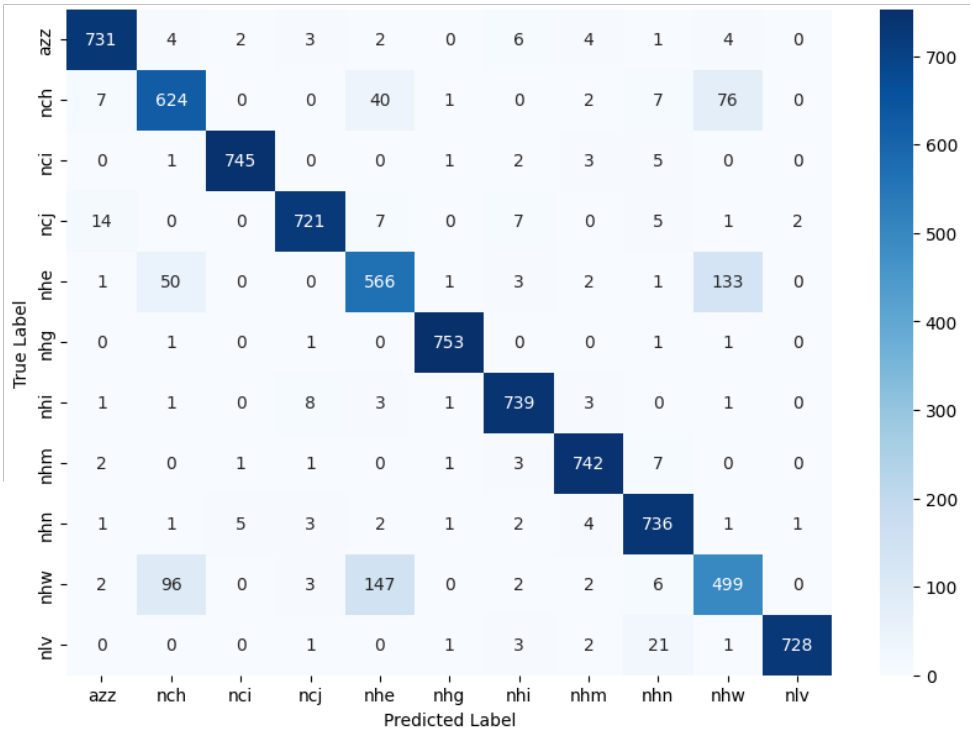


Figure 2: Confusion matrix of the ensemble model's predictions. The most frequently confused variants are the three Huasteca varieties (nhe, nhw, nch), which share a large number of phonological, morphological, and vocabulary features, and are typically considered members of a single Eastern Nahuatl subgroup Canger (2001).

| Lang. | LR | SVM | XLMR | Ens. |
|---|---|---|---|---|
| azz | 0.95 | **0.96** | 0.95 | **0.96** |
| nch | 0.79 | 0.78 | 0.78 | **0.82** |
| nci | 0.98 | 0.98 | 0.98 | 0.98 |
| ncj | 0.94 | 0.95 | 0.94 | **0.97** |
| nhe | 0.72 | 0.72 | 0.72 | **0.75** |
| nhg | 0.99 | 0.99 | 0.99 | 0.99 |
| nhi | 0.96 | 0.96 | 0.95 | **0.97** |
| nhm | 0.95 | 0.96 | 0.96 | **0.97** |
| nhn | 0.94 | 0.94 | 0.94 | **0.96** |
| nhw | 0.65 | 0.64 | 0.59 | **0.70** |
| nlv | **0.98** | **0.98** | 0.97 | **0.98** |
| **Avg** | **0.90** | **0.90** | **0.89** | **0.91** |

Table 5: Performance (F1 score) of the three component systems (Logistic Regression LR, Support Vector Machine SVM, and XLMR language-model) and the ensemble of the three.

| Class | Top 10 discriminative features (LR) |
|---|---|
| azz | ta, w, ne, j, k, eju, ten, no, de, ejua |
| nch | ta, t, cati, ti, j, hu, inta, tal, ten, tali |
| nci | y, uh, yn, in, z, au, auh, l, v, ll |
| ncj | in, n, ahmo, ahm, iya, huan, ehh, ehhu, ehhua, aki |
| nhe | ax, tlen, j, hu, len, queja, ueja, tl, eja, imo |
| nhg | ie, e, que, tie, hua, ua, nu, tlo, tli, o |
| nhi | h, u, ua, eh, queh, ueh, uan, aua, ou, iu |
| nhm | k, non, a, o, h, nin, z, i, ce, kej |
| nhn | k, w, l, wa, ce, in, ka, tl, ll, ts |
| nhw | catli, atli, catl, j, hu, ij, an, se, tli, quen |
| nlv | k, h, w, ki, eh, iwa, iw, iwan, ka, lli |

Table 6: Top 10 discriminative character $n$-grams per class from the logistic regression model trained on TF–IDF features.

recall (0.66), with persistent confusion against *nch* (approximately 15% of confusions), which supports a continuum interpretation for closely related regional languages rather than a sharply separable boundary. The complete confusion matrix heatmap can be seen in Figure 2.

### 5.1 Linguistic vs. Orthographic Features

The model appears to rely heavily on surface cues, including some well-established isoglosses as well as orthographic tendencies (artifacts of the specific texts in the corpus vs. dialectological features).

In Table 6, we show the top-most discriminative features in the Logistic Regression model. For example, one well-known isogloss among Nahuatl varieties is the "tl-t-l" distribution, where some varieties use "tl", others "t", and still others "l" for corresponding lexical items and morphemes. We see that this is valuable for distinguishing between varieties, since two "t-varieties" in our corpus, azz and nch, have sequences like *ta* and *ten* (*tla* and *tlen* in other varieties) in their top discriminative features. Other useful discriminative linguistic features are the lexical item *auh* for nci, a particle that is prevalent in Classical Nahuatl texts, and *ax*, a verbal negation prefix in the Huasteca region, as the top discriminative feature for nhe.

| Language | Precision | Recall | F1-score |
|---|---|---|---|
| azz | 0.96 | 0.94 | 0.95 |
| nch | 0.79 | 0.84 | 0.82 |
| nci | 0.97 | 0.97 | 0.97 |
| ncj | 0.95 | 0.95 | 0.95 |
| nhe | 0.72 | 0.77 | 0.75 |
| nhg | 0.99 | 1.00 | 0.99 |
| nhi | 0.95 | 0.94 | 0.94 |
| nhm | 0.94 | 0.94 | 0.94 |
| nhn | 0.93 | 0.91 | 0.92 |
| nhw | 0.70 | 0.65 | 0.67 |
| nlv | 0.98 | 0.96 | 0.97 |
| **Accuracy** | | | 0.89 |
| **Macro Avg** | 0.89 | 0.89 | 0.89 |

Table 7: Classification performance, final ensemble model, on orthographically normalized version of the balanced dataset. cf. rightmost column of Table 5.

Additionally, the list of top discriminative features highlights the value of orthographic cues in our experiment, such as *k* and *w* for nlv and nhn, substrings containing *y* for nci, and *h* and *u* (followed by a vowel) for nhi. It is important to note that, while some orthographic practices may be adopted by specific communities (e.g. the so-called "Tenango" orthography described in Pugh et al. (2025b)), orthographic patterns are largely an artifact of text/author rather than the Nahuatl variety. In order to evaluate the extent to which our model learned to perform LID via document-specific rather than language-specific features, we train and evaluate the ensemble model on an orthographically-normalized version of the corpus, produced using the py-elotl normalizer (Gutierrez-Vasques et al., 2025) set to the INALI standard. These results, shown in Table 7, show that indeed, to some extent, orthographic patterns alone contribute to some of the performance, since the normalized experiment shows small but consistent drops for all varieties.

### 5.2 The effect of domain

We also briefly explore the impact of the source text domain. Specifically, we compare performance on two small subcorpora (see Section 3.2) containing only the six languages where we have both Bible and non-Bible data. Table 8 shows results for Settings A (only non-Bible data) and B (mixed Bible and non-Bible data) for those six languages, in a very small data setting. Table 9 shows the same for five of those six languages, increasing the dataset size by removing the smallest language.

Although inclusion of Bible data doubles the amount of data available, performance in Setting B either stays the same as Setting A or decreases. This suggests that the Biblical register may introduce domain-specific noise that obscures authentic regional features. Variety-specific scores, though, fluctuate significantly with the inclusion of Bible data. For instance, in Table 8, the *nhw* variety drops from 0.84 to 0.77, while *nlv* increases from 0.82 to 0.91. This suggests that the model may be learning the standardized, formal voice of the Bible rather than capturing authentic regional cues. The domain of religious translations introduces a stylistic similarity that can distort performance metrics for specific variants, motivating our decision to prioritize naturally occurring text in our final model iterations.

| | Setting A (No Bible) | | Setting B (With Bible) | |
|---|---|---|---|---|
| **Variety** | **F1-Score** | **Support** | **F1-Score** | **Support** |
| azz | 0.95 | 74 | 0.90 | 147 |
| nch | 0.82 | 74 | 0.76 | 148 |
| nhi | 0.89 | 73 | 0.88 | 147 |
| nhm | 0.89 | 73 | 0.91 | 147 |
| nhw | 0.84 | 74 | 0.77 | 147 |
| nlv | 0.82 | 74 | 0.91 | 148 |
| **Macro Avg** | **0.87** | 442 | **0.86** | 884 |
| **Accuracy** | **0.87** | 442 | **0.86** | 884 |

Table 8: Performance comparison between non-Bible data (Setting A) and mixed data (Setting B), for 6 languages. Balanced settings, with 370 instances per language in Setting A and 735 in Setting B.

| Variety | Setting A (No Bible) F1-Score | Support | Setting B (With Bible) F1-Score | Support |
|---|---|---|---|---|
| azz | 0.93 | 162 | 0.95 | 325 |
| nch | 0.83 | 163 | 0.83 | 326 |
| nhi | 0.93 | 163 | 0.95 | 325 |
| nhm | 0.89 | 163 | 0.93 | 325 |
| nhw | 0.90 | 162 | 0.86 | 325 |
| **Macro Avg** | **0.90** | 813 | **0.90** | 1626 |
| **Accuracy** | **0.90** | 813 | **0.90** | 1626 |

Table 9: Performance comparison between non-Bible data (Setting A) and mixed data (Setting B), for 5 languages. Balanced setting, with 810 instances per language in Setting A and 1625 in Setting B.

## 6 Concluding remarks and future work

This paper introduces an ensembled text classification model to address the challenges posed by the lack of labeled corpus data for varieties of Nahuatl, a minoritized language family often treated incorrectly as a single language. By leveraging a weighted soft-voting ensemble combining Linear SVM (0.60), Logistic Regression (0.20) and XLM-R (0.20), the best system achieves a robust classification accuracy of 91% on naturally occurring test data.

Results indicate that transformers and character n-gram models can implicitly learn cross-variant correspondences robustly from naturally occurring text, effectively navigating the linguistic diversity of Nahuatl.

In future work, we plan to incorporate more conversational, social media, and community-generated content, to expand the scope of the system and to avoid formal and/or translated Bible data. We also plan to investigate the effectiveness of moving beyond the surface-level subwords we get from tokenization to a morphologically-motivated approach, which may capture useful morphological structures. This sort of approach may be especially useful for the agglutinative nature of Nahuatl languages.

Other future directions highlight the importance of engaging with speaker communities. First, we hope to integrate feedback from speaker communities to refine the classification system, with a particular focus on determining meaningful units of analysis. Finally, we plan to incorporate this classifier to develop variety-appropriate language processing tools such as spell-checkers, OCR post-correction, and machine translation.

## Limitations

This work has several limitations that should be considered when interpreting the results.

First, although we control for data imbalance through downsampling in our experimental design, there is a tradeoff between building balanced datasets and using all available data. Training on the pre-sampling corpus, which is uneven in both size and domain distribution, may improve performance for many varieties, but it may also bias the models toward higher-resource varieties and more represented genres.

Second, as shown in our orthographic normalization experiments, a portion of model performance appears to rely on surface-level orthographic conventions rather than deeper linguistic distinctions. This could potentially limit generalization to unseen writing styles or communities with different conventions. Some varieties in our corpus are heavily skewed toward a single domain; most notably Classical Nahuatl (nci), which appears exclusively in the poetry domain. This raises the possibility that the model may be learning domain or register signal rather than variety-specific linguistic features for these classes. Future work should evaluate performance on held-out domains to disentangle these effects.

Third, the inclusion of domain-specific data, particularly Biblical text, introduces stylistic regularities that may not reflect everyday language use, affecting the classifier's flexibility. Additionally, our sentence-level formulation may not fully capture broader discourse-level cues relevant for variety identification.

Fourth, the system does not perform open-set language identification: it will assign one of the 11 Nahuatl variety labels to any input, including non-Nahuatl text. In practice, this classifier should be composed with a general-purpose LID system that first identifies the input as Nahuatl.

Finally, given the limited availability of labeled data and focus on 11 varieties, the generalizability of our findings to the full diversity of Nahuatl variants remains an open question.

## Ethical considerations

This work uses data from an Indigenous language, where questions of data ownership, representation, and use are especially important. In line with CARE principles for indigenous data governance,

we recognize that not all data used in this study can or even should be freely redistributed.

Portions of the corpus were collected through direct collaboration with speakers and are used here with permission for research purposes only. Any future data release will be done in accordance with this to ensure that community preferences, access restrictions, and appropriate attributions are respected.

We also acknowledge that automatic language or variety identification systems may have unintended consequences if used without care. For example, misclassifications could affect downstream applications such as educational tools or language technologies, potentially reinforcing incorrect associations or privileging certain variants over others. Also, modeling decisions that treat language varieties as fixed and separable categories may not align with speakers' own linguistic identities or practices.

For these reasons, we emphasize that such systems should be developed and applied in collaboration with communities, with attention to their goals, expectations, and concerns.

## 7 Acknowledgements

Thanks to the anonymous reviewers for helpful and interesting suggestions. Thanks also to those who helped with gathering additional corpus data, without whom this project would have been even more challenging. Specifically, we extend our thanks to Rodrigo Ortega Acoltzi, who translated "The Birth of the Fifth Sun," Lydia Leija, who translated "Theft of Music," and Chicome Itzcuintli Amatlapalli, who authored both books. And thanks to members of the LECS Lab at CU Boulder for comments and suggestions. This work was supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another."

## References

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial Evaluation Campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. "If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages". In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China. Association for Computational Linguistics.

Hilda Berenice Aguayo Rousell and Juan Manuel Piña Osorio. 2016. Expressions of racism in a sample of university students in Mexico. *Sinéctica*, (46).

Una Canger. 1988. Subgrupos De Los Dialectos Nahuas (1988). In J. Kathryn Josserand and Karen Dakin, editors, *Smoke and Mist: Mesoamerican Studies in Memory of Thelma D. Sullivan.Part. Oxford: BAR International Series 402 (Ii)*, volume 402 of *BAR lnternational*, pages 473–98. BAR, Oxford.

Una Canger. 2001. Nahuatl dialectology: A survey and some suggestions. *Tonos: Revista de Estudios Filológicos*.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita C. Holbrook, Raymond Lovett, Simeon Materechera, Mark A. Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE Principles for Indigenous Data Governance. *Data Sci. J.*, 19:43.

Peter A. Chew, Steve J. Verzi, Travis L. Bauer, and Jonathan T. McClain. 2006. "Evaluation of the Bible as a Resource for Cross-Language Information Retrieval". In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 68–74, Sydney, Australia. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Çağrı Çöltekin and Taraka Rama. 2016. "Discriminating Similar Languages with Linear SVMs and Neural Networks". In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 15–24, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexis Conneau and Kartikay Khandelwal. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, Shruti Rijhwani, Katharina Von Der Wense, and Manuel Mager. 2025. "Findings of the AmericasNLP

2025 Shared Tasks on Machine Translation, Creation of Educational Material, and Translation Metrics for Indigenous Languages of the Americas". In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Lastra de Suárez et al. 1986. Las áreas dialectales del náhuatl moderno.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abteen Ebrahimi and Katharina Kann. 2021. "How to Adapt Your Pretrained Multilingual Model to 1600 Languages". In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

J.I.E. Farfan. 2019. *Nahuatl Contemporary Writing: Studying Convergence in the Absence of a Written Norm*. University of Sheffield.

Szymon Gruda, Gregory Haimovich, and John Sullivan. 2023. Lexical creativity in modern Nahuatl: An analysis of multidialectal data. *Lingua*, 285:103488.

María Ximena Gutiérrez-Vasques. 2018. EXTRACCIÓN LÉXICA BILINGÜE AUTOMÁTICA PARA LENGUAS DE BAJOS RECURSOS DIGITALES.

Ximena Gutierrez-Vasques, Robert Pugh, Victor Mijangos, Diego Barriga Martínez, Paul Aguilar, Mikel Segura, Paola Innes, Javier Santillan, Cynthia Montaño, and Francis Tyers. 2025. "Py-Elotl: A Python NLP package for the languages of Mexico". In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 38–47, Albuquerque, New Mexico. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016a. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016b. "Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Miguel Figueroa-Saavedra, Carlos-Emiliano González-Gallardo, Graham Ranger, and Martha Lorena-Avendaño-Garrido. 2026. Classifying several dialectal Nawatl varieties. *arXiv preprint arXiv:2601.02303*.

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Martha Lorena Avendaño Garrido, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Graham Ranger, Carlos-Emiliano González-Gallardo, Elvys Linhares-Pontes, Patricia Velázquez-Morales, and Luis-Gil Moreno-Jiménez. 2025. "$\pi$-YALLI : un nouveau corpus pour des modèles de langue nahuatl / Yankuik nawatlahtolkorpus pampa tlahtolmachiotl". In *Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, pages 802–816, Marseille, France. ATALA \\& ARIA.

Juan-José Guzman-Landa, Jesús Vázquez-Osorio, Juan-Manuel Torres-Moreno, Ligia Quintana Torres, Miguel Figueroa-Saavedra, Martha-Lorena Avendaño-Garrido, Graham Ranger, Patricia Velázquez-Morales, and Gerardo Sierra-Martínez. 2025. A symbolic algorithm for the unification of nawatl word spellings. In *Mexican International Conference on Artificial Intelligence*, pages 141–154. Springer.

Magnus Pharao Hansen. 2013. Nahuatl in the plural: Dialectology and activism in Mexico. In *Proceedings of the American Anthropological Association, Annual Meeting*.

INALI. 2009. *Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*. Instituto Nacional de Lenguas Indigenas, México, D.F.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Amanda Kann. 2024. "Massively Multilingual Token-Based Typology Using the Parallel Bible Corpus". In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11070–11079, Torino, Italia. ELRA and ICCL.

Terrence Kaufman. 2001. The history of the Nawa language group from the earliest times to the sixteenth century: Some initial results. *Paper posted online at http://www. albany. edu/anthro/maldp/Nawa. pdf. University of Pittsburgh*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System*

*Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yolanda Lastra. 1986. *Las areas dialectales del nahuatl moderno*. Universidad Nacional Autonoma de Mexico, Instituto de Investigaciones Antropologicas.

Eric Le Ferrand, Cian Mohamed Bashar Hauser, Joshua Hartshorne, and Emily Prud'hommeaux. 2025. "Faithful Transcription: Leveraging Bible Recordings to Improve ASR for Endangered Languages". In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 333–342, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Walter Lehmann. 1920. Die Sprachen Zentral-Amerikas in ihren Beziehungen zueinander sowie zu Sud-Amerika und Mexiko, 1/2. *Zentral-Amerika, Teil I*.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. "The Usefulness of Bibles in Low-Resource Machine Translation". In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 44–50, Online. Association for Computational Linguistics.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. "From Priest to Doctor: Domain Adaptation for Low-Resource Neural Machine Translation". In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.

Mike Maxwell. 2005. "Language Documentation: The Nahuatl Grammar". In *Computational Linguistics and Intelligent Text Processing*, pages 474–485, Berlin, Heidelberg. Springer Berlin Heidelberg.

Thomas Mayer and Michael Cysouw. 2014. "Creating a massively parallel Bible corpus". In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Luis Armando Mercado-Campos. 2023. Design and implementation of an NMT system for Spanish-Nahuatl. Master's thesis, Universidad del País Vasco / Euskal Herriko Unibertsitatea, Donostia-San Sebastián, España, June.

Garrett Nicolai and David Yarowsky. 2019. "Learning Morphosyntactic Analyzers from the Bible via Iterative Annotation Projection across 26 Languages". In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1765–1774, Florence, Italy. Association for Computational Linguistics.

Justyna Olko, R. Borges, and John Sullivan. 2018. Convergence as the driving force of typological change in Nahuatl. *STUF - Language Typology and Universals*, 71:467 – 507.

Claudia Parodi. 2011. Multiglosia virreinal novohispana: el náhuatl. *Cuadernos de la ALFAL*, 2:89–101.

Magnus Pharao Hansen. 2014. The East-West split in Nahuan Dialectology: Reviewing the Evidence and Consolidating the Grouping. In *Friends of Uto-Aztecan Workshop*.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. "Universal Dependencies for Western Sierra Puebla Nahuatl". In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.

Robert Pugh and Francis Tyers. 2021. Investigating variation in written forms of Nahuatl using character-based language models. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 21–27.

Robert Pugh and Francis Tyers. 2024a. Experiments in multi-variant natural language processing for Nahuatl. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 140–151, Mexico City, Mexico. Association for Computational Linguistics.

Robert Pugh, Francis Tyers, and Marivel Huerta Mendez. 2021. Towards and Open Source Finite-State Morphological Analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 80–85.

Robert Pugh, Francis Tyers, and Brian OConnor. 2025a. 8. Implementación de Identificación de Idiomas para las Lenguas Indígenas de México. *UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO*, page 88.

Robert Pugh and Francis M. Tyers. 2024b. A Universal Dependencies Treebank for Highland Puebla Nahuatl. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Robert Pugh, Cheyenne Wing, María Ximena Juárez Huerta, Ángeles Márquez Hernandez, and Francis Tyers. 2025b. "Ihquin tlahtouah in Tetelahtzincocah: An annotated, multi-purpose audio and text corpus of Western Sierra Puebla Nahuatl". In *Proceedings of the 2025 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. Association for Computational Linguistics.

Rico Sennrich and Alexandra Haddow. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Ana Tona, Guillaume Thomas, and Ewan Dunbar. 2023. "A morphological analyzer for Huasteca Nahuatl". In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 112–116, Remote. Association for Computational Linguistics.

Francis Tyers and Robert Pugh. 2023. "A finite-state morphological analyser for Highland Puebla Nahuatl". In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 103–108, Toronto, Canada. Association for Computational Linguistics.

Bhartendoo Vimal and S Anupama Kumar. 2020. Application of logistic regression in natural language processing. *Int J Eng Res*, 9(06).

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. "A Report on the DSL Shared Task 2014". In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. "Overview of the DSL Shared Task 2015". In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.

# A Appendix

## A.1 Corpus details

Table 10 shows in more detail the distribution of text types per language in our corpus.

Table 11 shows detailed information about language varieties and sources in the Axolotl corpus (Gutierrez-Vasques et al., 2016b,a), from which we use a subset.

| Lang | Source | # of Sentences | Avg. Sent. Length (tokens) |
|---|---|---|---|
| nhn | El Ajolote de Xochimilco (Story) | 133 | 3.04 |
| | Classes (Daily life) | 1,700 | 13.12 |
| | Historias para Niños (Stories) | 2,300 | 2.39 |
| nhi | UD (Transcripts) | 813 | 8.24 |
| | UD (Transcripts) | 90,200 | 5.48 |
| nhg | Conversaciones (Daily life) | 30 | 4.10 |
| | Diccionario de Frases (Daily life) | 87,800 | 4.10 |
| nch | UTexas (Educ.) | 1,000 | 2.30 |
| | Mitos en Náhuatl (Myths) | 104,800 | 5.33 |
| ncj | Bible (Religion) | 88,900 | 3.92 |
| nci | 4 Poemas Cortos (Poetry) | 121 | 3.26 |
| | Poemas en Náhuatl (Poetry) | 5,300 | 23.47 |
| nlv | La lengua de los Aztecas (Educ.) | 28 | 9.82 |
| | Guiones Bilingües (Legal) | 340 | 1.17 |
| | Bible (Religion) | 14,300 | 11.62 |
| azz | Conversaciones (Transcripts) | 2,900 | 6.31 |
| | OpenSLR (Mixed) | 40,900 | 8.66 |
| | Bible (Religion) | 127,300 | 4.53 |
| nhe | Conversaciones (Mixed) | 149 | 161.07 |
| | Bible (Religion) | 168,600 | 5.02 |
| nhm | Various Books (Mixed) | 1,900 | 10.26 |
| | Bible (Religion) | 5,600 | 7.46 |
| nhw | Constitución (Legal) | 1,400 | 3.71 |
| | Translated songs (Daily life) | 2,100 | 1.14 |
| | Bible (Religion) | 366,300 | 5.11 |

Table 10: Dataset Statistics: Variant, Source (Domain)

| Corpus | Sentences | Variant | Domain |
|---|---|---|---|
| Mitos y cuentos nahuas de la Sierra Madre Occidental | 93,466 | azz | Short stories |
| Los cuentos en náhuatl de Doña Luz Jiménez | 34,801 | azz | Short stories |
| Primer Axotli Libro | 15,794 | nci | Didactic |
| López Austin, Alfredo. Augurios y Abusiones | 32,392 | nci | Historical |
| Documentos nauas de la Ciudad de México del siglo XVI | 123,473 | nci | Historical |
| Testimonios de la Antigua Palabra | 43,444 | nci | Historical |
| Chimalpain Cuauhtlehuanitzi | 34,203 | nci | Historical |
| Historia de México narrada en náhuatl y español | 10,881 | nci | Historical |
| Anales de Tepeteopan | 8,162 | nci | Historical |
| Nican Mopohua | 9,176 | nci | Literature |
| Veinte Himnos Sacros de los Nahuas | 9,296 | nci | Literature |
| Trece Poetas del Mundo Azteca | 8,405 | nci | Literature |
| La tina negra y roja | 28,657 | nci | Literature |
| La llave del náhuatl | 26,957 | nci | Literature |
| La tierra nos escucha | 7,146 | nci | Literature |
| Teatro náhuatl II Selección y estudio Crítico | 38,939 | nci | Literature |
| Recetario Nahua el Norte de Veracruz | 24,040 | nci | Recipes |
| Recetario Nahua de Milpa Alta | 18,836 | nci | Recipes |
| Antología del cuento náhuatl | 36,469 | nci | Short stories |
| Adivinanzas | 326 | nci | Short stories |
| Lo que relatan de antes. Kuentos tének y nahuas de la Huasteca | 14,134 | nci | Short stories |
| Reyes García, Luis y Christensen, Dieter. El anillo de Tlalocan | 23,102 | nci | Short stories |
| Garibay, vida económica de Tenochtitlan | 30,794 | nhe | Historical |
| Revista La lengua y cultura Nahuatl | 33,800 | nhe | Magazine |
| Untitled | 46,868 | nhm | Historical |
| Yancuitlalpan, tradicion y discurso ritual | 16,527 | nhm | Short stories |
| El Náhuatl de Tetzcoco en la Actualidad | 32,825 | nhn | Didactic |
| Método autodidáctico español-náhuatl náhuatl-español | 29,166 | nhn | Didactic |
| La voz profunda | 6,778 | nhn | Literature |
| Revista Amerindia | 12,710 | nhn | Magazine |
| Cuéntos Indígenas de México | 496 | nhw | Musical |
| **TOTAL** | **851,563** | | |

Table 11: Axolotl Corpus Overview (Sorted by Variant and Domain)