A Universal Dependencies Treebank for Highland Puebla Nahuatl

Robert Pugh and **Francis Tyers**
Indiana University, Department of Linguistics
pughrob@iu.edu, ftyers@iu.edu

Abstract

We present a Universal Dependencies (UD) treebank for Highland Puebla Nahuatl[1]. The treebank is only the second such UD corpus for a Mexican language, and supplements an existing treebank for another Nahuatl variant. We describe the process of data collection, annotation decisions and interesting syntactic constructions, and discuss some similarities and differences between the Highland Puebla Nahuatl treebank and the existing Western Sierra Puebla Nahuatl treebank.

1 Introduction

Annotated linguistic corpora are an essential component of natural language processing (NLP), and are fundamental for training and evaluating models and applications. Furthermore, consistently-annotated corpora enable quantitative, comparative linguistic analyses.

The Universal Dependencies (UD) project (Nivre et al., 2020) is a widely-used annotation framework whose aim is to provide a consistent schema for representing morphological and dependency-based syntactic phenomena for all of the world's languages. Since a UD corpus contains rich information for all aspects of a standard NLP pipeline (tokenization, part-of-speech tagging, morphological analysis and syntactic parsing), it is well-suited as a resource for NLP application development. UD treebanks have also been successfully used in quantitative syntax research (Kiss and Thomas, 2019; Tyers and Henderson, 2021) and large, multilingual typological linguistic studies (Naranjo and Becker, 2018; Levshina, 2019).

The development of annotated linguistic corpora for endangered, indigenous, and/or marginalized languages is critically important to ensure their inclusion in the domain of digital language technology as well as to improve their representation in, and therefore the validity and robustness of, multilingual corpus-based studies.

In this paper, we present one such corpus, created using the UD framework, for Highland Puebla Nahuatl, a language spoken in central Mexico. This corpus will not only contribute to the linguistic study and NLP development for the Highland Puebla variant, but is also an important addition to the other existing Nahuatl treebank (for the Western Sierra Puebla variant) in that it can enable research into quantitative morphosyntactic dialectology and multi-dialectal NLP for Nahuatl.

2 Highland Puebla Nahuatl

Nahuatl is a polysynthetic, agglutinating Uto-Aztecan language continuum spoken throughout Mexico and Mesoamerica. Mexico's *Instituto Nacional de Lenguas Indígenas* (INALI) recognizes 30 distinct variants (INALI, 2009).

Highland Puebla Nahuatl, (or *Sierra Puebla Nahuatl*, also referred to by INALI as *Náhuatl del noreste central*, ISO-639-3 *azz*, henceforth HPN) is a Nahuatl variant group spoken by about 70,000 people (Ethnologue's 2007 estimate) in the Northeastern Sierra region of the state of Puebla, Mexico (see Figure 1) in 24 municipalities (INALI, 2009).

This particular Nahuatl variant has been the subject of documentary and descriptive linguistic efforts (Key, 1960; Robinson, 1970; Key and Key, 1953), and there are at least two published dictionaries (Key and Richie de Key, 1953; Cortez Ocotlán, 2017). Intergenerational transmission of HPN is declining, resulting in language shift. Along with all other indigenous languages in Mexico, HPN is considered at risk of being lost (INALI, 2012).

HPN is a member of the Eastern branch of Nahuatl (also commonly referred to as Aztec) along with Huasteca Nahuatl, Gulf Nahuatl, and Pipil/Nawat.

[1] https://github.com/UniversalDependencies/UD_Highland_Puebla_Nahuatl-ITML

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1393–1403

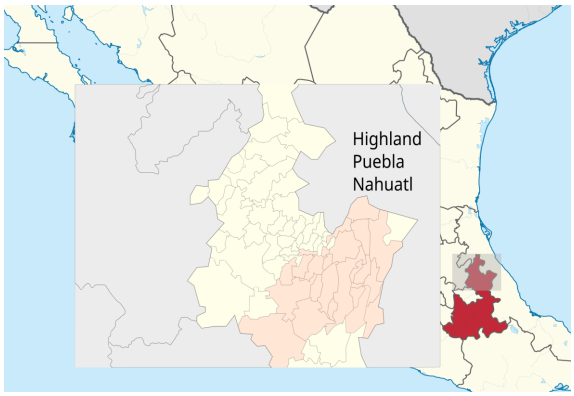


Figure 1: A map highlighting the 24 municipalities where HPN is spoken in the Sierra Norte de Puebla region of Mexico.

Its speakers are believed to have brought the language to the Sierra de Puebla region either during the first migration of Nahuatl-speakers from Northwestern Mexico (Canger, 1988), or as part of an early migration out of the valley of Mexico in around 800 C.E. (Kaufman, 2001).

In terms of isoglosses, HPN is a 't-dialect' since it has /t/ where central Nahuatl varieties typically have /tl/, has /i/, instead of /e/, corresponding to Proto-Uto-Aztecan **u*, and has a number of distinguishing lexical features (Lastra, 1986).

3 Related work

Despite UD's recent wide adoption in the computational linguistics community and over 100 UD treebanks for a diverse set of languages, treebanks for languages of the Americas are few: corpora of over 10,000 tokens for Mbyá Guaraní (Thomas, 2019), K'iche' (Tyers and Henderson, 2021), Western Sierra Puebla Nahuatl (WSPN) (Pugh et al., 2022), a slightly smaller (9,000 tokens) treebank for Guajajara, and a number of very small corpora (less than 2,000 tokens) for various languages of Brazil, e.g. Apurina (Rueter et al., 2021; Martín Rodríguez et al., 2022) and St. Lawrence Island Yupik (Park et al., 2021). Tyers et al. (2023) describe an early-stage, ongoing effort to annotate a treebank of Classical Nahuatl, the set of Nahuatl varieties represented in texts from the early colonial period.

Work on resources and NLP development for HPN includes a large audio corpus with transcriptions and translations (Amith et al., 2019) which has been used to train ASR and speech translation models (Shi et al., 2021). Pugh and Tyers (2023) present a finite-state morphological analyzer for the language, developed on a superset of the data used for the present treebank.

Research on the syntax of contemporary Nahuatl includes Flores Nájera (2019)'s treatment of the simple clause of the Nahuatl spoken in Tlaxcala, and explorations of non-configurationality and polysynthesis in Hueyapan Nahuatl (Pharao Hansen, 2010) and Western Sierra Puebla Nahuatl (Sasaki, 2021). Additional research in this area has focused on specific syntactic constructions such as relative clauses (de la Cruz Cruz, 2010; Flores Nájera, 2021; Pharao Hansen, 2015) and anti-passives (Flores Nájera, 2019). Additionally, a non-trivial amount has been written about Classical Nahuatl morphosyntax (Lockhart, 2001; Launey, 1994; Sasaki, 2018a, 2012), and the impact of contact on language change in Nahuatl over time (Hill et al., 1999; Canger and Jensen, 2007). We see the development of this and other morphosyntactically-annotated corpora for Nahuatl as a significant contribution to this area.

4 Corpus

The treebank data come from four sources:

1. The Axolotl Nahuatl-Spanish parallel corpus (Gutierrez-Vasques et al., 2016), which contains text in at least 5 Nahuatl variants, including the HPN variety. We sampled the texts in this variant for annotation.

2. Amith et al. (2019), containing many hours of recorded, transcribed, and translated monologues and dialogues discussing local plants, personal narratives, and stories. All of the sentences from this dataset have an accompanying audio file path, which we have included in the sentence metadata.

3. "Science outreach in indigenous languages" (*Divulgación de la ciencia en lenguas indígenas*), published by the Mexican Society of Physics (*Sociedad Mexicana de Física*, SMF) and INALI, translated into many indigenous Mexican languages, presenting scientific concepts for public consumption[2].

4. 32 example sentences from an introductory course in HPN offered in person in the municipality of Tetela de Ocampo, Puebla.

[2] https://site.inali.gob.mx/SMF/Libros2.0/nhtl/index.html

Table 4 breaks down these four sources with information about genre and token/sentence volume. Notably, both the Axolotl and the OpenSLR data largely contain descriptions of plants and their uses. Since these two sources make up most of the corpus, the treebank is biased toward this specific subject matter.

4.1 Orthography

Nahuatl has no singular accepted standard orthography, though a number of orthographic standards have been proposed (de la Cruz Cruz, 2014). As such, Nahuatl textual sources from different authors are very likely to differ in their orthography. The four data sources of our treebank are internally orthographically consistent. The texts taken from both the Axolotl corpus texts and the SMF use the orthographic standard associated with the Public Education Secretariat (*Secretaría de Educación Pública*, SEP), which uses 'j' for /h/, 'u' for /w/, and 'k' for /k/, and does not represent vowel length. The OpenSLR data generally follows the orthography proposed by INALI, which is similar except that it uses 'h' for /h/ and 'w' for /w/. Additionally, the OpenSLR data represents vowel length with a colon[3] and occasionally uses an apostrophe to represent abbreviated words that are not fully pronounced in speech (e.g. *t'titipitstoti* is an abbreviated form of *tiktitipitstoti* 'you will be blowing it'). In the treebank, we keep the original orthography in the token forms (with the exception of vowel length), and include a normalized form in the 10th ("MISC") column, designated for including other miscellaneous information not covered by the standard UD fields. For lemmas, we use an orthography that was taught in the Nahuatl course for adult learners given in the municipality of Tetela de Ocampo, Puebla (TO) in the summer of 2022. This broadly follows the SEP, but with the addition of the letter *h* which is used before *u* for /w/ after vowels or at the beginning of words. For example SEP *ueueyi*, TO *huehueyi* 'big', SEP *mochiua*, TO *mochihua* "it is made". Importantly, conversion between Nahuatl orthographies is relatively straightforward, and tools have been developed for doing this automatically[4].

[3]Representing vowel length is not common in contemporary Nahuatl orthographies, and we remove it from the word forms in the corpus.

[4]https://github.com/ElotlMX/py-elotl

5 Annotation decisions

Automatic processing was used to put the source text into CONLL-U format. The lemma, part-of-speech, and morphological analysis were all initially performed automatically using a finite-state morphological analyzer for the language. The second author manually reviewed the automated results, and annotated the syntactic trees for all of the sentences using the UD Annotatrix annotation tool (Tyers et al., 2017). Then, the first author reviewed all of the trees and any disagreements were discussed, with the differing views argued for, until both came to an agreement. The trees were updated with Arborator Grew (Guibon et al., 2020). This approach to annotation was chosen because, at the moment, there is no well-defined set of guidelines for Nahuatl UD parsing. Thus the creation of this treebank, in conjunction with prior work on WSPN, serves to further define and document annotation guidelines.

In the following sections, we describe in more detail the processes and decisions made during annotation, highlighting a few noteworthy constructions in the treebank.

5.1 Morphology

Lemmas, universal part-of-speech (UPOS) tags, and morphological features were generated using the morphological analyzer described in Pugh and Tyers (2023), using a table to map the Apertium morphological tags to the equivalent UD feature-value pairs. The results were then verified manually. Lemmas are given in a single orthographic norm, regardless of the orthography of the surface form. We generally follow the Western Sierra Puebla Nahuatl treebank (Pugh et al., 2022) with respect to morphological features, e.g. `NounType=Relat` for Relational Nouns and layered `Number` and `Person` features for subjects and object of verbs, and for possessors of possessed nouns.

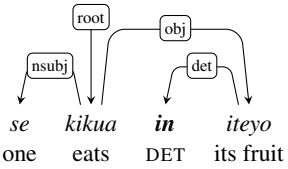
5.2 Syntactic constructions

"in": determiner, pronoun, and subordinator: In Classical Nahuatl linguistics, the word *in* has historically been analyzed as a subordinator or adjunctor that introduces subordinate (often nominal) predicates (Andrews, 1975; Launey, 1994), though in contemporary variants its distribution is much more similar to a determiner (Sasaki, 2018b). In the HPN treebank, we see "in" play multiple roles: determiner (Example 1), pronoun (Example 2), and

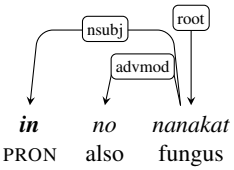
Source	Genre	Trees	Tokens
Gutierrez-Vasques et al. (2016)	nonfiction	660	5,002
Amith et al. (2019)	spoken	499	3,882
Sociedad Mexicana de Física	nonfiction	68	1,088
Pedagogical examples	grammar	33	116
Totals		1,261	10,088

Table 1: A summary of data sources for the treebank. The genres are consistent with (Müller-Eberstein et al., 2021). Two of the four sources, both of which mainly deal with botanical discussions, account for most of the data.

subordinator (Example 3).

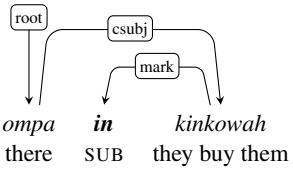


(1) "Its fruit is eaten."



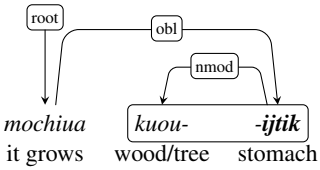
(2) "This is also a mushroom/fungus."

In the Western Sierra Nahuatl treebank, it was noted that "in" as a subordinator was commonly used in focus constructions. These cases are also present in the HPN data, such as in Example 3.



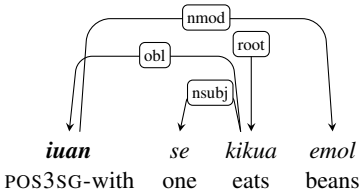
(3) "It is there (where) they buy them."

Relational nouns and Adpositions Nahuatl, like other Mesoamerican languages, typically uses a subclass of Noun, called Relational Nouns (RNs), to express spatial relations, where other languages might use an Adposition[5]. Example 4 shows a compounding form with the RN *ijtik*, "inside" (literally "stomach"). In Example 5, we see the RN *uan*, the comitative "with," with a third-person singular possessive prefix corresponding to its complement, *emol* "bean sauce." This example also illustrates how a RN and its complement need not be contiguous, resulting in a non-projective tree.



(4) "It grows in the woods."

When an RN occurs compounded with another noun, we tokenize the compound noun into two words and represent their syntactic relationship so that it parallels the case of the possessed RN (with an `obl` relation from the head to the RN, and an `nmod` relation from the RN to the nominal complement). For an illustration, compare Examples 5 and 4.



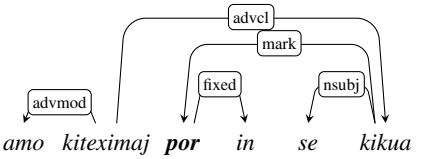
(5) "It is eaten with beans."

The most frequent RNs in the corpus are *-tech* "in/on", *-ka* the instrumental "with", *-huan* the comitative "with", and *-tsintan* "beneath."

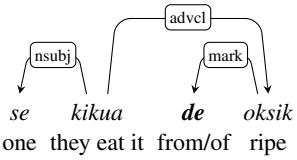
Extended contact (over 500 years) with Spanish has resulted in the borrowing of some prepositions. Our corpus includes examples of 6 prepositions: *de* 'from/of', *por* 'for', *kemej* from Spanish *como* "like/as", *para* "for/in order to", *hasta* "until", and *a* "to". In many cases, these prepositions behave similarly in Nahuatl as they do in Spanish. Some prepositions have also seen changes or extensions in their usage in Nahuatl. For example, *por* has become part of a fixed subordinating expression *por in* meaning "because" (Example 6). The diversity of the borrowed preposition *de* has been

[5]In fact, some Nahuatl linguists and educators refer to RNs as Postpositions (e.g. Campbell and Karttunen (1989)) due to the fact that they can be compounded with Nouns. We prefer analyzing them as Nouns, since many of them are clearly etymologically Nouns, and they can take possessive morphology.

well-documented (Hill et al., 1999). In Example 7 from the corpus, it used to introduce an adverbial clause.

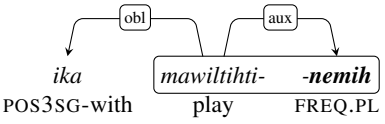


(6) "They don't cut it because it is eaten (its fruit)."



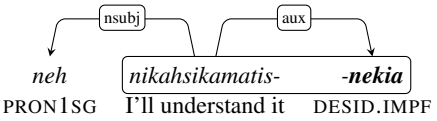
(7) "It is eaten when it is ripe."

Verb-auxiliary compounds Nahuatl has a set of aspectual auxiliaries, derived from verbs of motion, that compound with verbs and take tense/aspect suffixes. We tokenize these compound forms, adding an aux relation from the main verb. For most of these forms, the main verb has the so-called "ligature" *-ti-* appended to it (as in Example 8).



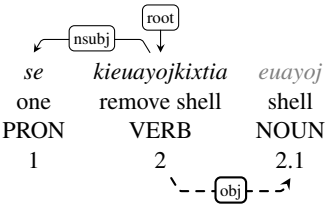
(8) "Children go around playing with it."

The auxiliary *neki*, "to want", is unique in that, instead of using the ligature, it compounds with a main verb in the future tense, with the auxiliary taking tense/aspect suffixing, as in Example 9.



(9) "I wanted to understand it."

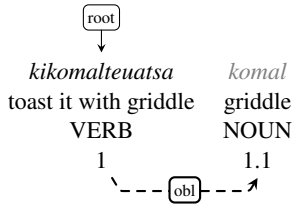
Noun incorporation Nahuatl can typically incorporate nominal arguments into the verb, a process common to polysynthetic languages. In the HPN data, we see both object incorporation (Example 10) and oblique incorporation (Example 11).



(10) "Its shell/skin is removed (lit. one shell-removes it[6])."

To annotate noun incorporation, it is important to represent the incorporated argument in the treebank even though it is not realized as a separate word. We follow Tyers and Mishchenkova (2020)'s recommendation, also used in Pugh et al. (2022), using UD's "enhanced" dependency layer to represent the syntactic relationship between the verb and its incorporated argument.

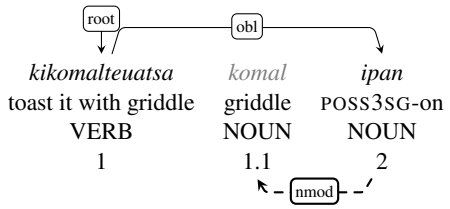
In these cases, the token listed with a decimal token id (e.g. 2.1) is not a separate token in the sentence. Rather, we create an empty node for it in the "enhanced dependency" layer to represent its syntactic relationship to the sentence even though it is incorporated into the verb.



(11) "they(sg.) toast it on the griddle."

An interesting fact about oblique incorporation like that in Example 11 is that, under our analysis (described in Section 5.2), the incorporated noun is not dependent on the verb, since in a version of this sentence without noun incorporation we would expect the oblique to be an RN and the noun, in this case *komal* "griddle", would be its nmod complement.

Furthermore, Example 12 shows a sentence not found in our corpus but attested in HPN, wherein the head of the incorporated oblique noun is present in the sentence.



(12) "They(sg.) toast it on the griddle."

There is a related phenomenon documented in West Greenlandic (Malouf, 2000) where dependents of an incorporated noun can show up external to the verb. In the Nahuatl case, however, the verb-internal incorporated noun is dependent on a word external to the verb.

[6]The verb *kixtia* is causative. As such,

A detailed exploration of this syntactic phenomena is beyond the scope of this paper, and we designate it for future work. Our hope is that annotating the syntactic relations of incorporated nouns in our treebank will make the analysis of such syntactic phenomena more accessible.

6 Comparing two Nahuatl treebanks

In this section, we take stock of the two existing Nahuatl treebanks, of the Highland Puebla (HPN) (the present paper) and Western Sierra Puebla (WSPN) (Pugh et al., 2022) variants, comparing some basic linguistic properties and metadata in order to better inform future work that uses these corpora for comparative analysis or NLP.

Genre: The two Nahuatl treebanks have quite different genre profiles: The HPN treebank presented here breaks down into essentially half spoken monologue/dialogue discussing plants and plant uses, and the other half non-fiction written text about the same topic. The WSPN treebank is more diverse with respect to its genre distribution, with substantial portions of the text containing fiction, spoken personal narratives, and elicited example sentences.

Number of tokens, types, and trees: The two treebanks have nearly the same number of tokens, 10,356 and 10,088, respectively, though the HPN treebank achieves this number of tokens with far more trees (about 300 more), with a shorter average sentence length. This is expected given the nature of the data sources for each treebank described above. Figure 2 shows the difference in sentence length distribution between the treebanks. The type-token ratios of the two treebanks, 0.23 for WSPN and 0.18 for HPN, also suggests greater lexical diversity in the WSPN treebank.

Parts of speech: The distribution of UPOS tags in the two treebanks is similar, with nouns and verbs by far the most frequent part-of-speech and NUM and X the least frequent (the latter used typically only with the dependent in a `goeswith` relation). At the high-frequency end, `NOUN` and `VERB` occur with nearly identical frequency, whereas in the WSPN treebank, `VERB` is by far the most frequent tag. `AUX` also has a much higher frequency in the WSPN treebank (10% vs. 1.5% in the HPN treebank), a fact likely due to an isoglossic difference between the two variants: WSPN prepends

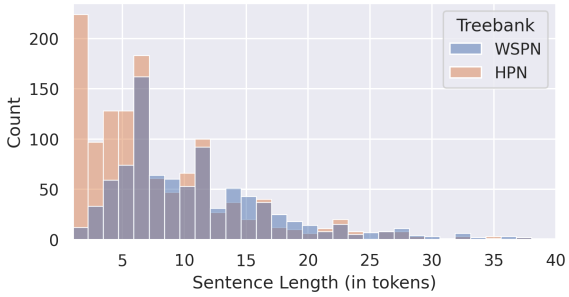
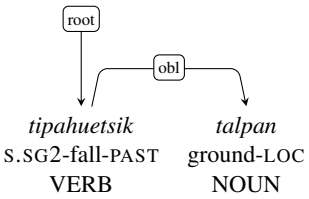


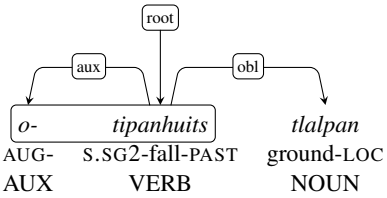
Figure 2: A visualization of the difference in sentence-length (ignoring punctuation) distribution between the Highland Puebla Nahuatl (HPN) and Western Sierra Puebla Nahuatl (WSPN) treebanks. The HPN treebank has a much higher concentration of short sentences.

the "antecessive" *o-* on all verbs in the past. Compare the two annotations of the phrase "you fell on the ground" in HPN (13) and WSPN (14).

(13)



(14)



Since this word is annotated as `AUX` in the treebank, the category becomes much more frequent.

Other morphosyntactic features: In order to get a sense of the extent to which linguistic and genre differences might affect the corpora, we examine the distribution of some morphosyntactic properties in both treebanks. Figure 4 shows the distributions of person/number of verbal subjects. While the general ranking is the same (3rd-person > 1st-person > 2nd-person) and (sg > pl), the HPN treebank is more heavily skewed, with nearly 70% of verbal subject being 3rd-person singular. The WSPN treebank has relatively more cases of 1st and 2nd-person subjects, which is likely due to the presence of narratives and stories.

We also examine the distribution of dependency labels in the two treebanks (Figure 5). The differ-

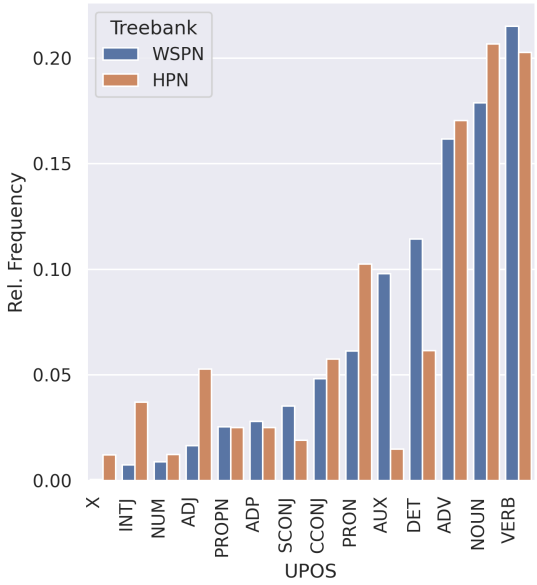


Figure 3: A visualization of the part-of-speech (UPOS) distributions in the two treebanks. The large differences for the INTJ and ADJ categories can likely be explained by the HPN treebank having more spoken content (leading to more INTJ) and containing plant descriptions (leading to more ADJ).

ence in frequency of the aux relation is also due to a dialectal difference wherein WSPN requires an "antecessive" element *o-* before verbs in the past tense. The difference in the frequency of det is not entirely clear, though it may also be due to dialectal differences (usage of the determiner *in* is variable across Nahuatl variants, and may simply be used more or less depending on the variant or speech community). Finally, the discrepancy in advmod frequency may be due to genre/subject matter differences (more descriptive sentences resulting in more adverbs), dialectal differences (the HPN treebank has a number of instances of the adverbial *yek* 'well' as a separate word that adverbially modifies a verb or adjective, whereas in WSPN it is frequently incorporated as an adverbial prefix), and/or differing annotation standards between the treebanks (e.g. the word *ok*, "still", is treated as an auxiliary in WSPN, but as an adverb in HPN[7]).

We have shown how, for a number of linguistic features, the two Nahuatl treebanks pattern similarly, yet for others, dialectal and genre differences result in different linguistic profiles between them. These types of treebank heuristics may be useful

[7]As a result of this work and subsequent further analysis of the WSPN treebank, the latter corpus will be updated in this respect for the next release, and will more closely match the HPN treebank.

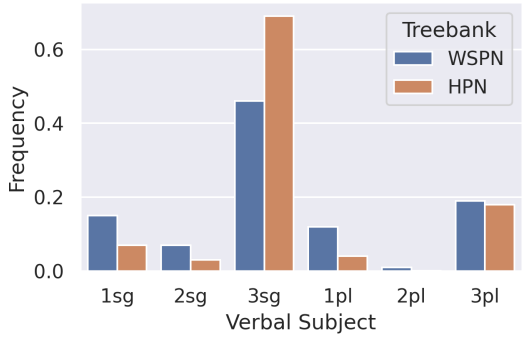


Figure 4: Comparing the frequencies of verbal subject person/number in the two treebanks. The higher frequencies of 1st and 2nd person subjects in WSPN is likely reflective of genre.

for cross- or multi-dialectal NLP systems.

7 Training a Highland Puebla Nahuatl Parser

In addition to corpus-based linguistic analyses, one of the principal uses of an anntotated treebank is the for the development of an NLP pipeline. To accompany the above discussion of the linguistic resource itself, here we present the results of training a neural model on the treebank data.

We use the MaChAmp toolkit (van der Goot et al., 2021) to fine-tune contextual embeddings from the pretrained multilingual BERT (mBERT) model[8] on lemmatization (Lemma), part-of-speech tagging, morphological analysis, and dependency parsing. The model leverages multi-task learning, such that all of the tasks share encoder parameters, but each has its own unique decoder: a transformation-rule classifier (Straka, 2018) for lemmatization, a softmax layer on the contextual embeddings for part-of-speech tagging and morphological analysis, and a deep biaffine parser for dependency parsing (Gardner et al., 2018). For all tasks we use the default hyperparameters.

We trained the model for 100 epochs, selecting for each fold the epoch with the best performance on the evaluation set, where performance was defined as the sum of the accuracies of each task. Thus, we could potentially achieve better results by selecting the best performing model for each task independently. Since the goal in the present

[8]We use the bert-base-multilingual-cased model, and a total of 183M parameters (most of these are from the mBERT model, not the task-specific decoders). Experiments were run on a Quadro RTX 6000, and training took approximately 3 hours.

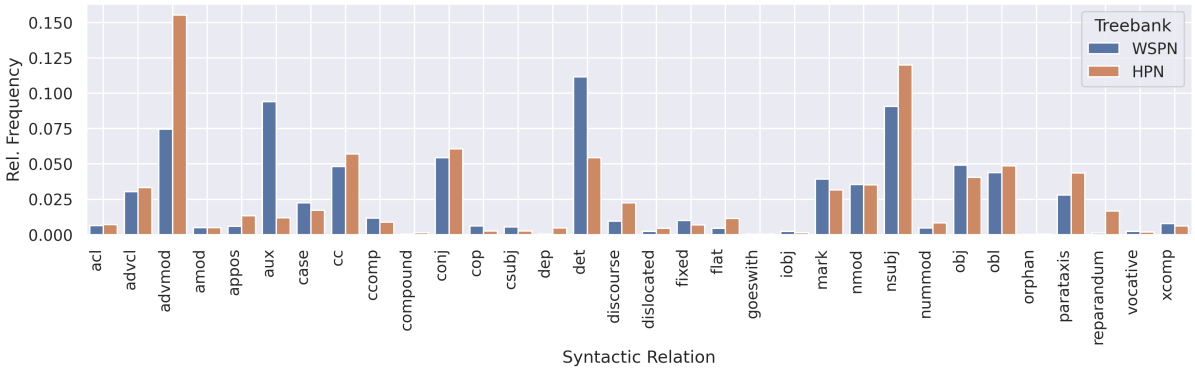


Figure 5: A comparison of the distribution of dependency labels between the HPN and WSPN treebanks. Many of the dependency labels have approximately the same relative frequency in the two treebanks, with a few exceptions: The discrepancy in the frequency of `aux` is, as discussed in the text, due to the WSPN variant's use of the antecessive auxiliary *o-*, a dialectal feature not present in HPN. The differences in `advmod` and `det` frequencies can be attributed to a combination of dialectal difference, genre, and potentially annotation inconsistencies.

state of the project is to benchmark UD parsing for this new treebank, we leave optimizations and improvements on Nahuatl dependency parsing as an avenue of future research.

The results, shown in the "In domain" column of Table 2, are promising particularly given the low training data volume. The model performs best on the part-of-speech tagging task, where it is on-par with the average performance of state-of-the-art UD parsing systems on all UD treebanks with training set (Straka et al., 2019). This is not surprising, since Nahuatl has a number of affixes unique to a given part of speech (e.g. tense/aspect suffixes for verbs, nominal plural and diminutive suffixing on nouns, etc.). Lemmatization is the worst-performing task outside of dependency parsing, likely due to Nahuatl's complex morphology and the fact that the lemmas are sometimes in a different orthography than the forms, so the lemmatization process also implicitly learns orthographic conversion. About 58% of the tokens in the corpus have lemmas identical to their forms, so an accuracy of nearly 0.9 is a significant improvement over a majority baseline.

	Result	
Metric	*In domain*	*Out domain*
Lemmas	89.8 ± 1.1	—
UPOS	94.5 ± 0.8	87.6
Morpho Feats	91.7 ± 1.2	—
UAS	79.8 ± 2.2	86.7
LAS	72.7 ± 2.0	76.6

Table 2: Accuracy of a neural, multi-task UD parsing system trained on the HPN treebank. 'In domain' is the average accuracy (and standard deviation) when performing 10-fold cross-validation. 'Out domain' is the performance on a sample of text not included in the treebank, with different subject material (ritual practices). For the out-of-domain evaluation, we skip lemmatization and morphological features due to time constraints.

7.1 Performance out of domain

Given the noted homogeneity of subject matter in our corpus, we hypothesized that the our model will likely perform substantially worse on text that differs in genre or subject.

In order to begin testing this hypothesis, we sample 200 sentences of an HPN text from a different domain and evaluate predictions of the model trained on our treebank[9]. The data also comes from the Axolotl corpus, but is from a book dealing with ritual practice in the Sierra Norte region. We selected 200 random sentences from this source and ran them through the model described in the previous section. We then manually corrected the trees via Arborator-Grew, and compared the predicted trees with the corrected trees. Due to time constraints, we did not manually correct lemmas or morphological features, and do not evaluate them.

Our results are listed in the "Out domain" column in Table 2. Predictably, UPOS tagging drops substantially compared to in-domain data. However, counter-intuitively, we see better dependency parsing performance on the out-of-domain data.

[9]We use one of the models trained during 10-fold cross validation, so it is in fact only trained on 9/10 of the treebank.

On closer inspection, this can be understood by more closely examining the data: Nearly all of the sentences in the out-of-domain sample are between 1 and 5 words long, with a non-trivial number of 1- and 2-word sentences. Given the relative syntactic simplicity in this evaluation data, the higher parsing performance is unsurprising.

8 Concluding remarks

We have presented a UD treebank for Highland Puebla Nahuatl, which accompanies a treebank for a different Nahuatl variety as one of two UD treebanks for indigenous languages of Mexico (and, incidentally, one of only two UD treebanks for Nahuatl languages). We observe a number of similarities with the existing Nahuatl treebank, as well as some noteworthy novel syntactic constructions. Additionally, we trained a neural model on the treebank, demonstrating the ability to build a relatively high-performing NLP pipeline for the language.

We leave detailed experimentation of the NLP pipelines for Nahuatl as future work. Furthermore, we hope that this treebank in conjunction with the other Nahuatl treebank enables quantitative exploration of syntax within the domain of Nahuatl dialectology.

Acknowledgements

We would like to thank Patricia Aguilar Romero, don Pedro Rivera, and Elesban Landero Berriozábal for their help with the work described in this manuscript. In addition we would like to thank the anonymous reviewers for their helpful comments.

References

Jonathan D. Amith, Amelia Dominguez Alcántara, Hermelindo Salazar Osollo, Ceferino Salgado Castañeda, and Eleuterio Gorostiza Salazar. 2019. Audio corpus of Sierra Nororiental and Sierra Norte de Puebla Nahuat(l) with accompanying time-code transcriptions in ELAN.

J.R. Andrews. 1975. *Introduction to Classical Nahuatl*, 2nd edition. University of Texas Press.

Joe R. Campbell and Frances Karttunen. 1989. Foundation course in Nahuatl grammar.

Una Canger. 1988. Nahuatl dialectology: A survey and some suggestions. *International Journal of American Linguistics*, 54(1):28–72.

Una Canger and Anne Jensen. 2007. Grammatical borrowing in nahuatl. *EMPIRICAL APPROACHES TO LANGUAGE TYPOLOGY*, 38:403.

Pedro Cortez Ocotlán. 2017. *Diccionario Nahuat–Español de la Sierra Nororiental del Estado de Puebla*. Tetsijtsilin, Tzinacapan, Cuetzalan.

Victoriano de la Cruz Cruz. 2010. Las cláusulas relativas en el náhuatl de Teposteco, Chicontepec, Veracruz.

Victoriano de la Cruz Cruz. 2014. La escritura náhuatl y los procesos de su revitalización. *Contribution in New World Archaeology*, 7:187–197.

Lucero Flores Nájera. 2021. Headless relative clauses in Tlaxcala Náhuatl. In *Headless Relative Clauses in Mesoamerican Languages*, pages 79–110. Oxford University Press.

Lucero Flores Nájera. 2019. *La gramática de la cláusula simple en el náhuatl de Tlaxcala*. Ph.D. thesis, Centro de Investigactiones y Estudios Superiores en Antropología Social.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. 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.

J. H. Hill, K. C. Hill, J. Farfán, and G. L. Cruz. 1999. *Hablando mexicano : la dinámica de una lengua sincrética en el centro de México*.

INALI. 2009. *Catalogo De Las Lenguas Indigenas Nacionales: Variantes Linguisticas De Mexico Con Sus Autodenominaciones Y Referencias Geoestadisticas*. Instituto Nacional de Lenguas Indigenas, México, D.F.

INALI. 2012. *México: Lenguas indígenas nacionales en riesgo de desaparición*. Instituto Nacional de Lenguas Indigenas, México.

Terrence Kaufman. 2001. The history of the nawa language group from the earliest times to the sixteenth century: some initial results.

Harold Key. 1960. Stem construction and affixation of Sierra Nahuat verbs. *International Journal of American Linguistics*, 28(2):130–145.

Harold Key and Mary Richie de Key. 1953. *Vocabulario Mejicano de la Sierra de Zacapoaxtla, Puebla*. Instituto Lingüístico de Verano, México, D.F.

Mary Key and Harold Key. 1953. The phonemes of sierra nahuat. *International Journal of American Linguistics*, 19(1):53–56.

Angelika Kiss and Guillaume Thomas. 2019. Word order variation in Mbyá Guaraní. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 121–129.

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

Michel Launey. 1994. *Une grammaire omniprédicative*. CNRS-Editions.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.

James Lockhart. 2001. *Nahuatl as written: Lessons in older written Nahuatl, with copious examples and texts*. Stanford University Press.

Robert Malouf. 2000. West greenlandic noun incorporation in a monohierarchical theory of grammar.

Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon, and Fabrício F. Gerardi. 2022. Tupían language ressources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58, Marseille, France. European Language Resources Association.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. How universal is genre in Universal Dependencies? Technical Report arXiv:2112.04971, arXiv.

Matías Guzmán Naranjo and Laura Becker. 2018. Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, 155, pages 91–104. Linköping University Electronic Press.

J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.

Hyunji Hayley Park, Lane Schwartz, and Francis Tyers. 2021. Expanding Universal Dependencies for polysynthetic languages: A case of St. Lawrence Island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142, Online. Association for Computational Linguistics.

Magnus Pharao Hansen. 2010. Polysynthesis in Hueyapan Nahuatl: The status of noun phrases, basic word order, and other concerns. *Anthropological linguistics*, pages 274–299.

Magnus Pharao Hansen. 2015. Dialectal variation in contemporary Nahuatl relative clause formation. AILS Seminar.

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. 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.

Dow F. Robinson. 1970. *Aztec studies 2: Sierra Nahuat word structure*. Summer Institute of Linguistics.

Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen, and Niko Partanen. 2021. Apurinã Universal Dependencies treebank. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 28–33, Online. Association for Computational Linguistics.

Mitsuya Sasaki. 2012. R-marking: Referential person affixes in Classical Nahuatl nouns. Master's thesis, University of Tokyo.

Mitsuya Sasaki. 2018a. Cityless air makes free: Characteristics of free variation in modern Nahuatl. [Draft].

Mitsuya Sasaki. 2018b. In predecible: Hacer tangible la sintaxis nahua. In *Seminarios de Lenguas Indígenas, UNAM*.

Mitsuya Sasaki. 2021. *Configurationality in Ixquihuacan Nahuatl*. Ph.D. thesis, University of Tokyo.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63, Online. Association for Computational Linguistics.

Milan Straka. 2018. Udpipe 2.0 prototype at CoNLL 2018 ud shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Milan Straka, Jana Straková, and Jan Hajic. 2019. Evaluating contextualized embeddings on 54 languages in pos tagging, lemmatization and dependency parsing. *ArXiv*, abs/1908.07448.

Guillaume Thomas. 2019. Universal dependencies for Mbyá Guaraní. In *Proceedings of the third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77.

Francis Tyers and Robert Henderson. 2021. A corpus of K'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20.

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204.

Francis Tyers, Robert Pugh, and Valery Berthoud F. 2023. Codex to corpus: Exploring annotation and processing for an open and extensible machine-readable edition of the florentine codex. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 19–29, Toronto, Canada. Association for Computational Linguistics.

Francis Tyers, Mariya Sheyanova, and Jonathan Washington. 2017. UD Annotatrix: An annotation tool for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.