Universal Dependencies for Western Sierra Puebla Nahuatl

Robert Pugh[†]**, Marivel Huerta Mendez**[∗]**, Mitsuya Sasaki**[∗]**, Francis M. Tyers**[†]

† Indiana University
∗ Independent
{pughrob, ftyers}@iu.edu
{marivelhm1, mitchara}@gmail.com

Abstract

We present a morpho-syntactically-annotated corpus of Western Sierra Puebla Nahuatl that conforms to the annotation guidelines of the Universal Dependencies project. We describe the sources of the texts that make up the corpus, the annotation process, and important annotation decisions made throughout the development of the corpus. As the first indigenous language of Mexico to be added to the Universal Dependencies project, this corpus offers a good opportunity to test and more clearly define annotation guidelines for the Mesoamerican linguistic area, spontaneous and elicited spoken data, and code-switching.

Keywords: nahuatl, universal dependencies, treebank, syntax, morphology

1. Introduction

Linguistically-annotated corpora are critical resources for natural language processing and computational linguistics. Statistical models for virtually all tasks in these areas, including word- and sentence-tokenization, morphological segmentation and analysis, and syntactic parsing, are commonly trained using collections of annotated text. Rules-based systems, as well, can leverage annotated corpora as ground-truth for performance evaluation or as a reference for rule development. The Universal Dependencies (UD) project[1] (Nivre et al., 2016) is a widely-used annotation framework whose aim is to provide a consistent schema for morphological and syntactic phenomena for all of the world's languages. An annotated UD corpus contains rich information for all aspects of a standard NLP pipeline including tokenization, part-of-speech tagging, and morphological and syntactic analysis, making it a highly valuable resource for the development of NLP applications. Since the annotation schema is intended to be language-independent, the resulting annotated corpora can (and in fact are intended to) be leveraged for multilingual NLP systems and cross-lingual transfer for downstream tasks. In descriptive linguistics research, questions about syntactic patterns and tendencies in a language can be approached quantitatively with a large enough corpus (Kiss and Thomas, 2019; Tyers and Henderson, 2021). UD corpora are also useful for large-scale, multilingual corpus analysis (Naranjo and Becker, 2018; Levshina, 2019).

Given UD's goal of achieving consistent, cross-linguistically viable annotation guidelines for all of the world's languages, it is crucial to prove out the existing guidelines on a diverse set of languages and domains. While the set of UD treebanks represents an impressive level of linguistic diversity, there are still a number of language families and linguistic areas that have yet to be analyzed with UD. Particularly relevant to our work, indigenous languages of Latin America are under-represented in UD both in terms of the number of languages within a treebank, and in terms of the sizes of the existing treebanks. With respect to the genre in existing treebanks, the overwhelming majority of the datasets represent written language, much of which is edited and polished (Müller-Eberstein et al., 2021).

In the remainder of this paper, we describe in detail the development of an annotated corpus of Western Sierra Puebla Nahuatl, an indigenous language variant spoken in central Mexico. In so doing, we offer an example of how the existing UD guidelines can be applied to largely-spoken and frequently code-switched[2] datasets for a morphologically-rich Mesoamerican language. Section 2 provides some background about the language. In Section 3, we give a cursory overview of recent work in computational language technology for indigenous languages of the Americas, efforts to include such languages in the UD project, and the existing descriptive work on Nahuatl syntax. Section 4 describes the texts that make up our corpus. In section 5, we discuss in detail the annotation process and decisions made, focusing on tokenization, lemmatization, part-of-speech tagging, and syntactic constructions. Finally, Section 6 reports the results of a baseline tagger/parser for the new treebank.

2. Western Sierra Puebla Nahuatl

Nahuatl is a polysynthetic and agglutinative language continuum spoken throughout Mexico and Mesoamerica, belonging to the Nahuan branch of the Uto-Aztecan language family. Western Sierra Puebla Nahu-

---

[1] http://www.universaldependencies.org

[2] Throughout the paper, we use the term "code-switching" and "code-mixing" to refer to any obvious mixing of Nahuatl and Spanish words. Admittedly, the boundary between code-swtiching and borrowing can at times be fuzzy, particularly given that some Spanish words have been in common use in Nahuatl for nearly five centuries. However, an in-depth discussion of language contact in the Nahuatl context is out of scope for the present paper.
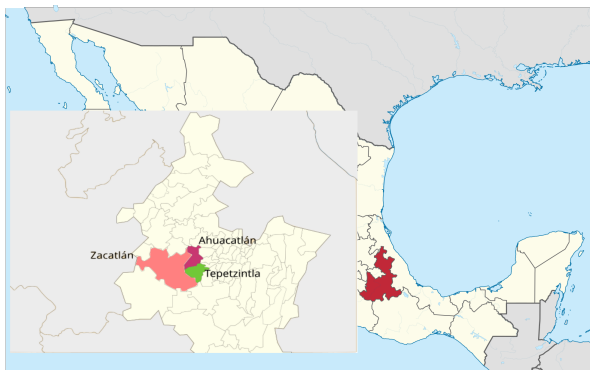
Figure 1: A map showing the location of the three municipalities from which the texts of our corpus originate: Zacatlán, Ahuacatlán, and Tepetzintla.

atl (*Náhuatl de la Sierra Oeste de Puebla*, alternatively Zacatlán-Ahuacatlán-Tepetzintla Nahuatl, ISO-639: `nhi`) is one of the 30 officially recognized Nahuatl variants (INALI, 2009), spoken in the Northwestern Sierra region of the state of Puebla, Mexico, primarily in the municipalities of Zacatlán, Ahuacatlán, and Tepetzintla. As of 2007, about 17,100 of the approximately 1.5 million Nahuatl speakers speak Western Sierra Puebla Nahuatl.

Nahuatl is one of the most widely-researched indigenous languages of the Americas, with a large body of linguistic research on both colonial varieties (including the so-called "Classical Nahuatl") (Carochi, 2001; Andrews, 1975; Lockhart, 2001; Launey and Mackay, 2011), and a number of contemporary variants (Langacker, 1977; Langacker, 1979; Hill et al., 1999; Flores Nájera, 2019). The Western Sierra Puebla Nahuatl variant, by contrast, has only recently been the subject of descriptive linguistic research (Sasaki, 2015). Petra Schroeder released an unpublished partial grammar and some descriptive work of the Western Sierra Puebla Nahuatl variety spoken in the town of San Miguel Tenango, Zacatlán (Schroeder and Tuggy, 2010; Schroeder, 2014; Schroeder, 2015). Mitsuya Sasaki published a sketch of the Western Sierra Puebla Nahuatl variant spoken in Ixquihuacan, Ahuacatlán (Sasaki, 2014), as well as dialectological overview of the Northern Sierra region (Sasaki, 2015) and an in-depth exploration of the question of non-configurationality in the language (Sasaki, 2021). Many Western Sierra Puebla Nahuatl speakers today also speak Spanish. Economic pressures, migration, and educational language policy have led to rapid language shift towards Spanish in most if not all Nahuatl-speaking communities (Olko and Sullivan, 2015).

3. Related Work

Research focused on computational resources and applications for indigenous languages of the Americas has recently grown in prevalence in the natural language processing community. The year 2021 saw the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (Mager et al., 2021b) and the second Congress of NLP for Indigenous Languages[3]. Computational work on Nahuatl has focused mainly on machine translation (Bello García et al., 2021; Gutierrez-Vasques et al., 2016; Mager et al., 2021a; Gutierrez-Vasques, 2015) and morphological segmentation and analysis (Farfan, 2019; Pugh et al., 2021; Kann et al., 2018; Eskander et al., 2019).

In terms of annotated corpora for Indigenous languages of the Americas, the Universal Dependencies project includes corpora of over 10,000 tokens for Mbyá Guaraní (Thomas, 2019) and K'iche' (Tyers and Henderson, 2021), and a number of very small corpora (less than 2,000 tokens) for various languages of Brazil (Akuntsu, Apurina, Guajajara, Kaapor, Karo, Makurap, Munduruku, and Tupinamba).

The most extensive descriptive work on the syntax of contemporary Nahuatl variants are Flores Nájera (2019)'s treatment of the simple clause of the Nahuatl spoken in Tlaxcala, and Sasaki (2021)'s analysis of word order and non-configurationality in Western Sierra Puebla Nahuatl. 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).

4. Corpus

We annotated texts from 4 sources, comprising a total of 10,356 tokens and 939 trees (see Table 4 for a breakdown and more details). The source texts represent Western Sierra Nahuatl spoken/written in each of the three municipalities where the language is prominently spoken, Zacatlán, Ahuacatlán, and Tepetzintla.

4.1. Orthographic and Regional Variation

For as long as Nahuatl has been written using the Latin alphabet, numerous written norms have been proposed and utilized (de la Cruz Cruz, 2014). Despite containing only text from a single Nahuatl variant, in a relatively small geographic region, and all written within the last 20 years, our corpus exhibits a large amount of orthographic variation (Table 2), both between and within sources. Such variation, particularly when dealing with a relatively small number of examples, can result in poor parser performance. While on the one hand, the simplest solution to orthographic variation is to normalize all sentences to a single uniform written standard, on the other hand this solution would lose information about decisions made by the authors as well as potentially important data about orthographic tendencies. In order to both be faithful to the original texts

---

[3]Congreso Internacional de Procesamiento de Lenguaje Natural para Lenguas Indígenas, Universidad Michoacana de San Nicolás de Hidalgo. `http://148.216.17.40/pln-wp/`

Source	Location	Genre	Tokens	Trees
Sasaki (2021)	Ixquihuacan	grammar/spoken	4,368	385
Schroeder (2014)	San Miguel	grammar	1,423	191
Pugh et al. (2021)	Omitlán	spoken	1,139	89
Márquez Hernández (2001)	San Miguel	non-fiction	431	30
Márquez Hernández (2003)	San Miguel	fiction	503	49
Márquez Hernández and Hernández Juárez (2005a)	San Miguel	fiction	252	20
Márquez Hernández and Hernández Juárez (2005b)	San Miguel	fiction	211	20
Márquez Hernández (2005a)	San Miguel	fiction	665	56
Márquez Hernández (2005c)	San Miguel	fiction	308	24
Márquez Hernández (2005b)	San Miguel	fiction	413	35
Márquez Pérez (2007)	San Miguel	fiction	643	40
Totals			10,356	939

Table 1: A summary of the data sources that make up the treebank corpus. The corpus represents Western Sierra Puebla Nahuatl varieties from Ixquihuacan (Ahuacatlán), San Miguel Tenango (Zacatlán) and Omitlán (Tepetzintla). The genres listed are consistent with (Müller-Eberstein et al., 2021).

Phoneme	Orthographic representations
/s/	*s, z, c*
/k/	*k, qu, c*
/w/	*w, u, hu*
/h/	*h, j*
/ts/	*ts, tz*

Table 2: An example of phonemes with multiple potential orthographic representations. Typically, the different orthographic representations correspond to specific orthographic norms, though in some cases there is variation within a single text or even sentence, e.g. *o<u>hc</u>on ame<u>jk</u>an [ohkon amehkan] 'They are like that.'*

and make the data more useful for training NLP tools, we include the original orthography in the 'lemma' column, and a normalized version of both the lemma and the surface form in the 'Misc' column (column 10), the latter using the recommended orthography of Mexico's National Institute of Indigenous Languages (INALI) (INALI, 2018)[4]. We performed the orthographic normalization using the `py-elotl` Python package[5]. We also observe regional variation between the texts. While the three municipalities from which our texts are collected all speak variants belonging to the same dialectal group, there are a number of regional distinctions that, though they might no pose much problem to speaker intelligibility, can in fact be challenging for automated parsers. For example, we observe two of three possible variations of the root meaning 'wood/tree': *koh-*, *poh-*, and *boh-*, which result in a number of varied forms for derived words, e.g. *kohtla/pohtla/bohtla* 'wooded area', *kowitl/powitl/bowitl* 'tree'. The data from San Miguel Tenango, Zacatlán also exhibit varia-

tion in the subject prefixes (*ni-*, *ti-*, etc.) and the third-person singular object prefix *ki-*, metathesizing in certain contexts (viz. *in-*, *it-*, *ik-*, etc.)[6].

4.2. Language Mixing

Given Nahuatl's centuries-long contact with Spanish and the high levels of bilingualism with Spanish among Western Sierra Puebla Nahuatl speakers, unsurprisingly our corpus contains extensive borrowing and code-switching between Nahuatl and Spanish. This includes Nahuatl affixes on a loaned Spanish root (Example 1), the use of just a few Spanish words in a sentence, frequently prepositions and conjunctions (Example 2), and majority-Spanish sentences. We include language information in the treebank to facilitate future work focused on bilingualism and code-switching. In the following glosses, language is written on the second line.

(1) *mo-tareas*
    nhi-spa
    POSS2SG-homework-PL
    'Your assignments'

(2) *pobre de notlawikal*
    spa   spa nhi
    poor of POSS1SG-husband
    'My poor husband'

5. Annotation Process and Decisions

We automatically converted text versions of the source files into CONLLU format using a Python script which also pre-populated fields that could be analyzed largely deterministically (e.g. adding POS tags for closed classes). Where translations were not provided, the

---

[4]All examples provided in this paper also conform to the same INALI written standard.

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

---

[6]These affixes are perhaps better understood as single consonants, like *n-*, with the main difference between the varieties having to do with where the *i*-epenthesis takes place. For a detailed discussion of this phenomenon, see (Schroeder and Tuggy, 2010)

second author, a native speaker of Western Sierra Puebla Nahuatl, first translated the sentences into Spanish. We then annotated them using UD Annotatrix (Tyers et al., 2017).

5.1. Tokenization

The first step in the annotation process, after sentence tokenization and CONLLU-formatting, is to identify the distinct syntactic (vs. orthographic) words in each sentence. Our texts reflect a tendency in Nahuatl writing to write auxiliary or adverbial words and clitics as agglutinated to their corresponding head (typically a verb).

By far the most common of these is the *augment o-*, an indicator that the corresponding verb describes an action that took place in the past. In Colonial Nahuatl, there are also strong syntactic reasons to consider the augment as a separate word (Launey and Mackay, 2011), though the evidence is not as strong for Western Sierra Puebla Nahuatl. Nonetheless, we separate it for consistency with other analyses of Nahuatl syntax and given the low cost of doing so, since it is trivial to re-attach it to the Verb if desired.

Other syntactic words that are commonly written agglutinated to other words in our corpus include the adverbial clitic *ya* 'already', which combines with the augment and verb (Example 3, the determiner *n*[7], and the optative particle *ma/mo* (the latter occurring in texts from the Tepetzintla).

(3) **yos**iyaw
already-PERF-S3SG-get.tired
"(S)he got tired."

(4) **ma**kilpih   *in* *itskwintli*
OPT-S3SG-O3SG-tie.up-PAST DET dog
"...that he tie up the dog"

In addition, given the frequent use of code-mixing with Spanish in our corpus, we follow tokenization decisions in the existing Spanish UD corpora, such as separating contractions (*del* →*de el*) and contracted object clitics (*dámelo* →*da me lo*).

5.2. Lemmatization

Nahuatl words are frequently the result of numerous derivational morphological processes, such as reduplication, applicative and causative suffixation, noun incorporation, and compounding. We leave all derivational morphemes in the lemma forms, stripping only inflectional morphology.

5.3. Part-of-Speech Tags

The Universal Part-of-Speech tag set defines a large set of word-classes, only a subset of which is typically used for a given language. Below, we provide a brief discussion of the major parts of speech we used for our corpus, and their motivations.

- **VERB**: Verbs are easily distinguishable from other word classes due to their inflectional and derivational morphology. They obligatorily inflect for person and number of subject in intransitives (5), and for subject and object in transitives (6)[8]. Tense and aspect are also marked on the verb. Derivational verbal morphology includes adverbials, reflexive, directionality, compounding and the incorporation of core arguments (typically objects).

  (5) *neh ni-ya-s*   *Zacatlán*
  SG1 S1SG-go-FUT Zacatlán
  'I will go to Zacatlán.'

  (6) *se-ki-yek-tlali-s*
  IMPERS-O3SG-WELL-put-FUT
  'We will fix it.'[9]

- **NOUN**: Most nouns take one of a small set of Absolutive endings in the singular, unpossessed form (*-tl*, *-tli*, *-li*) and can be inflected for number and diminution. There are generally two distinct plural suffixes depending on whether the noun is possessed (*-wan* if possessed, *-meh* otherwise[10]). Noun endings change depending on whether the noun is possessed (Compare Examples 7 and 8).

  (7) a. *wehxolo-tl*   b. *wehxolo-meh*
       turkey-ABS   turkey-PL
       'Turkey.'   'Turkeys.'

  (8) a. *to-wehxoloh*
       POSS1PL-turkey
       'Our turkey.'

      b. *to-wehxolo-wan*
       POSS1PL-turkey-PL
       'Our turkeys.'

  They can also act as predicates, taking subject prefixes, but are distinguished from verbs in that they cannot take tense morphology. Instead, to mark a predicative noun for tense, a copula *katki* is used.

---

[7]Historically, this is the subordinator *in*. In Western Sierra Puebla Nahuatl, it is often realized as *n* and written (and pronounced) together with the following word, particularly when that word begins with a vowel.

[8]In some cases, in transitive verbs with a third-person singular object, the object prefix can be omitted.

[9]*se-*, historically the impersonal subject prefix, typically takes the meaning of the 1st-person plural subject, but takes singular tense and aspect morphology.

[10]Historically Nahuatl has a number of different pluralization strategies, but these have largely reduced to *-meh* in the Western Sierra variant. Occasional idiosyncratic plural forms may also be found, e.g. via reduplication: *konetl* 'baby' →*kokoneh* 'babies'.

(9) a. *neh*   *ni-telpoka-tl*
PRON1SG S1SG-boy-ABS
'I am a boy.'

    b. *o-ni-katka*   *ni-telpoka-tl*
AUG-S1SG-cop S1SG-boy-ABS
'I was a boy.'

We also use the NOUN tag for the closed set of Relational Nouns, discussed in more detail in Section 5.5.

- **ADJ**: Adjectives in Nahuatl generally modify nouns (Example 10), and are usually derived from Nouns or Verbs. They can take nominal morphology (e.g. subject prefix, diminutive suffix) and can act as predicates to indicate the state of having some quality (Example 11).

(10) *se* **weyi** *kali*
one big house
'A big house.'

(11) *akmo*   *ni-**pitsawak***
no.more S1SG-skinny
'I am no longer skinny.'

- **ADV** Adverbs in Nahuatl can describe the manner, place, or time of the action taken by a verb. Some examples include *satepan* 'then', *mostla* 'tomorrow', and *nochipa* 'always', *ompa* 'there', and *nikan* 'here', among many others. Spanish adverbs, such as *allá* 'there', *entonces* 'then', and *ahora* 'now', are also common. The Spanish adverb *ya* 'already', in addition to being used in code-switching, has also become cliticized when combined with the perfective augment *o-* and written together.

A number of lexical items are tagged as both ADJ and ADV depending on whether they modify nouns or verbs, respectively.

(12) a. *chikawak in*   *kowatl*
ADJ   DET NOUN
strong   the snake
'The snake is strong.'

    b. *tsahtsi chikawak*
VERB ADV
yell strong
'She/he/it yells loudly.'

- **DET** Determiners always precede a noun, and include *in* (frequently written *n* and joined to the following noun), demonstratives *nin/ninkeh* 'this/these' and *non/nonkeh* 'that/those', and two other demonstrative determiners that are compounds of the words 'here' and 'there' with the

	Singular	Plural
1st person	nehwatl (neh)	tehwan
2nd person	tehwatl (teh)	namehwan
3rd person	yehwatl (yeh)	yehwan

Table 3: Personal pronouns in Western Sierra Puebla Nahuatl. The form in parentheses after the singular pronouns indicates short forms. Omitted from this table is the honorific 2nd-person singular pronoun, which varies between *tehwatsin* and *towatsin*. In Omitlán, Tepetzintla, the 2nd person plural pronoun is realized as *nimehwan* instead of *namehwan*.

clitic copula: *nikan-ka* and *ne-ka*. The latter two words can and do also occur as pronouns ('this/that thing') and verb phrases ("It is here/there"). Other common determiners are quantifiers such as *nochi* 'all' and *siki* 'some'.

(13) a. **ne-ka**   *tlaxcal o-k-wikak*
there-be tortilla AUG-O3SG-took
'He took that tortilla.'

    b. **ne-ka**   *mo-tlaxcal*
there-be POSS2SG-tortilla
'There is your tortilla.'

    c. *tleno* **ne-ka**
what there-be
'What is that?'

- **PRON** The personal pronouns in Western Sierra Puebla Nahuatl are displayed in Table 3. In addition to these and the demonstrative pronouns just mentioned, the frequent pronouns include interrogative pronouns *akin* 'who' and *tlen* 'what' and quantifiers *nochi* 'all' and *siki* 'some', which also frequently appear as determiners.

- **AUX** The UD guidelines define *auxiliaries* as any word that contributes tense, aspect, mood, or evidentiality to its verb head. The set of Nahuatl auxiliaries in our corpus includes:

  – *o*, known also as 'the augment', a clitic used to indicate a past action.

  – *ok*, indicating continual ongoing aspect ("still doing X"). This word typically occurs immediately to the left of the verb, but in some cases can follow it as well.

  – *mach*, an evidential particle indicating citation.

  – *mo/ma*, the optative particle, which is optional for second-person subjects (since there is also a distinct optative subject prefix for this case), but required otherwise. In most recorded variants it is realized as *ma*, but also appears as *mo* in our Tepetzintla data.

The verb *katki* "to be" is used in our corpus as a copula (see Example 9b), as well as to mean "there is/are". We tag it as AUX in the former case, and VERB in the latter.

- **SCONJ** There are a number of subordinating conjunctions used in our corpus, including *ijkwak* 'then', *nik* 'because', and *tla* 'if'. Additionally, the Spanish *porque* 'because', *que* 'that', and *hasta* 'until'. The subordinating conjunction *in* is used almost exclusively in "n-focalization" constructions, which we discuss in more detail in section 5.5.

- **CCONJ** Apart from *wan* 'and', our corpus predominantly uses Spanish loanwords such as *y* 'and' and *o* 'or', as coordinating conjunctions. Interestingly, the word *mas* 'but' also appears as a coordinating conjunction. This word was borrowed sometime during the colonial period, and has since become significantly less frequent in Spanish, while maintaining quite common usage in Nahuatl.

5.4. Morphological features

We seeded the generation of morphological features with a morphological analyzer for Western Sierra Nahuatl (Pugh et al., 2021). As the analyzer outputs sequences of morphological tags and not feature-value pairs, as required by UD, we transformed its output using a maximum-overlap algorithm to match the UD feature set, adding missing features manually.

5.5. Syntactic Relations

Core and non-core arguments Subjects of intransitive verbs, and both subjects and objects of transitive verbs, are obligatorily marked on the verbal via prefixation, and the phrase that is indexed by the agreement marker is often omitted (Example 14, and 15; optional independent arguments are in parentheses).

(14) (*neh*)   *o-ni-k-ilnamik*
     PRON3SG AUG-S1SG-O3SG-remembered
     'I remembered it.'

(15) *o-ni-k-ita-k*   (*se itskwintli*)
     AUG-S1SG-O3SG-see-PST one dog
     'I saw it (a dog)'

In ditransitive verbs, only the subject and indirect object prefixes can appear, unless the direct object has the 3rd person and is plural (Example 17.

(16) *ni-mits-ilwi-s   se historia*
     S1SG-I2SG-tell-FUT one story
     'I will tell you a story.'

(17) *o-Ø-tech-in-maka-ya*
     AUG-S3SG-O1PL-IO3PL-give-impf
     'She used to give them to us.'

Non-core arguments are never marked in the verb. They include relational nouns, or can be introduced by a Spanish preposition (example 18).

(18) *In   omitl Ø-ki-kui-h   para in*
     DET bone S3SG-O3SG-use-PL for   DET
     guitarras
     guitars
     'They use the bone for the guitars.'

The status of the determiner *in/n* The word *in* has been a topic of much interest to Nahuatl linguists over the years. Historically, it has been analyzed as a subordinator or adjunctor by e.g. Andrews (1975), though its usage has shifted in contemporary Nahuatl variants (and quite likely is not consistent across variants). We follow Sasaki (2018)'s analysis of *in* as a determiner in Western Sierra Puebla Nahuatl, as well as its continued use as a subordinating conjunction in clefting constructions, called "n-focalization" in Sasaki (2021).

Focalization and clefting Both core and non-core arguments can be focalized when followed by an *n*-marked subordinate clause (Example 19). This construction is likely historically related to the clefting-construction in colonial Nahuatl, and is referred to as n-focalization in Sasaki (2021). This type of clefting is also common in the Irish UD treebank (Lynn and Foster, 2016), and we analyze them similarly, with the focalized element as the matrix copular-sentence, and the subordinated clause as its clausal subject (csubj).

(19)
```
        ┌──advmod──┐ ┌─root─┐ ┌──csubj──┐
        │    ┌─det─┐│ │      │ │  ┌─mark─┐
        ▼    ▼      ▼│ ▼      ▼ │  ▼      ▼
       san   se   tamal      n     okikwah
       just  one  tamal     SUB   She ate it
       ADV   DET  NOUN     SCONJ   VERB
```
     "It is only one tamal that she ate."

Also common in the n-focalization constructions is the use of the third-person singular pronoun *yej* as a pronominal copula in the matrix clause, as shown in Example 20.

(20)
```
        ┌──cop──┐ ┌─root─┐ ┌──csubj──┐
        │  ┌─det─┐│ │     │ │  ┌─mark─┐
        ▼  ▼     ▼│ ▼     ▼ │  ▼      ▼
       ye   n    Juan      n     kichiwa
       COP DET  Juan      SUB   makes it
       PRON DET PROPN    SCONJ   VERB
```
     "It is Juan that makes it."

Relational Nouns As is common in the Mesoamerican linguistic area, Western Sierra Puebla Nahuatl uses "Relational Nouns" (RNs) to express the relation (typically) between a nominal and a predicate. These are typically similar in meaning to prepositions 'on', 'inside of', 'next to', 'with', etc. in English. They take the nominal morphology agreeing with the possessor, as in 21.

(21) *namech-tlali-s-keh* **i-pan**   *mitla-tl*
     O2PL-put-FUT-PL P3SG-on metate-ABS

"They will put you on top of the metate."

(22) *ni-k-niki*   *ni-mawilti-s*   **mo-wan**
S1SG-O3SG-want S1SG-playFUT P2SG-with

"I want to play with you."

While some prefer treating RNs as adpositions, e.g. Schroeder (2014) for Western Sierra Puebla Nahuatl, we analyze them as NOUN throughout the corpus. Our decision is primarily motivated by the fact that RNs take the possessive nominal morphology, and that, unlike with apositions, the related noun — analogous to the nominal complement in an adposition — is optional. For instance, example (21) above could just as easily appear as in (23). This latter motivation explains why, even for RNs that behave more like adpositions (e.g. their possessive morphology doesn't inflect to match the number of the possessor), we continue to treat them as a special subclass of nouns.

(23) *namech-tlali-s-keh*   **i-pan**
O2PL-put-FUT-SPL P3SG-on

"They will put you on top of it."

This is the same structure used for the genitive/possessive construction, as in (25).

(24) *i-tikak-wan*   *no-papa*
P3SG-shoe-PL P1SG-father

"My father's shoes."

(25) *i-tikak-wan*
P3SG-shoe-PL

"His shoes."

We annotate the relational noun with the `obl` relation, and `nmod` for the possessor when it is present. This maintains consistency with the UD annotation schema for K'iche' (Tyers and Henderson, 2021).

(26)
namech-tlali-s-keh   **i-pan**   *mitla-tl*
O2PL-put-FUT-PL P3SG-on metate-ABS
VERB   NOUN   NOUN
"They will put you all on the metate."

Relative clauses   Relative clauses in Western Sierra Puebla Nahuatl are typically (though not always) introduced with a relative pronoun (*tlen* 'what/that', *non* 'that', *akin* 'who') or relative adverb ('kampa', 'keman'). We annotate these cases using the `acl` and `nsubj` or `obj` relation, as in example 27.

(27)
tochtli   *tlen*   *simi*   *nakasmahmatok*
rabbit   PRON   very   ear-stand.up-DUR
NOUN   PRON   ADV   VERB
"rabbit who is always perking up its ears."

Annotating code-mixing   As we discussed in section 4.2, code-mixing is quite frequent in the corpus. We make efforts to ensure that, where applicable, our annotation decisions are consistent with those in Spanish treebanks. Code-mixing repetition, or the repetition of the same word or phrase in two different languages, is also present, particularly in the spoken data. We analyze this phenomenon with the `reparandum` relation, as in example 28.

(28)
ich   *n*   *ipapeles*   *itich*   *n*   *iamahwan*
on   DET   his papers   on   DET   his papers
nhi   nhi   nhi-spa   nhi   nhi   nhi-nhi

Given the frequent use of Spanish in the corpus, it is important that we maintain consistency with the existing Spanish UD treebanks.

Conjunction and parataxis   Throughout the corpus, we frequently observe sentences containing two or more adjacent clauses without an explicit coordinating conjunction. These cases are somewhat ambiguous with respect to their syntactic relationship, as they could be analyzed as conjunction (`conj`), which typically requires an explicit conjunction but can appear without one, e.g. in a list of items, or `parataxis`. Since the difference between these two relations can be ambiguous, we established a straightforward rule to facilitate ease-of-annotation: When the two clauses share an argument, use `conj`; otherwise, use `parataxis`.

(29)
otiyayah   *otitlapiyayah*   *ich*
we used to go   we used to herd   in
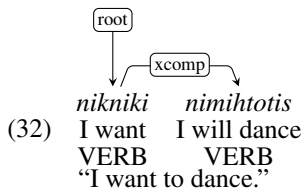VERB   VERB   NOUN
"We used to go (and) herd sheep in ..."

(30)
amo   *kwale*   *amo*   *oyahke*
NEG   good   NE   they went
ADV   ADJ   ADV   VERB
"It isn't good they didn't go."

In accordance with UD guidelines, we also use `parataxis` for reported speech.
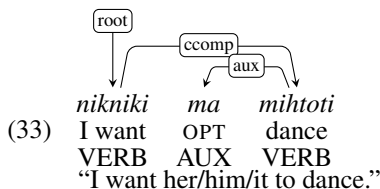
(31)
okilwuih: *¡ximewa!*
she said   get up!
VERB   VERB
"She said: get up!"

Clausal complements   A number of Nahuatl verbs can take clausal complements. We follow the UD guidelines in distinguishing between clausal complements with or without obligatory control. Verbs that do not control their clausal complements include *ita* 'see'

mati 'know', and *ilwia* 'tell', among others. These take the `ccomp` relation. Verbs such as *niki* 'want' and *pewa* 'begin', control the subject of their clausal complement, and are annotated with `xcomp`.

(32)

```
        ┌──[root]
        │   ┌──[xcomp]──┐
        ↓   ↓           ↓
     nikniki       nimihtotis
     I want        I will dance
     VERB          VERB
```
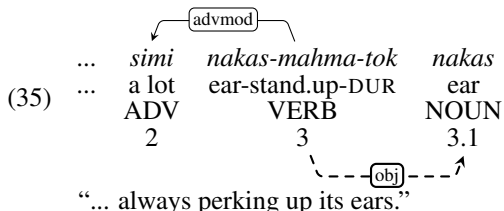"I want to dance."

For *niki* in particular, there are cases in which the subordinated clause has a different subject than its parent, but these cases require an optative marker, either the optative subject prefix *xi-* or the optative auxiliary *ma/mo*, and the form of the second verb, which must be in the future tense when its subject is controlled, changes. In this case the relation `ccomp` is used.

(33)

```
        ┌──[root]
        │   ┌────[ccomp]──────┐
        │   │     ┌─[aux]─┐   │
        ↓   ↓     ↓       ↓
     nikniki    ma      mihtoti
     I want     OPT     dance
     VERB       AUX     VERB
```
"I want her/him/it to dance."

Incorporation The incorporation of core argument nouns into the verb, as in example (34) is a well-known phenomenon common in polysynthetic languages (Sapir, 1911; Mithun, 1984).

(34) *o-mo-nacas-mahman*
AUG-REFL-ears-stand.up
'He perked his ears.'

However, the UD guidelines have not yet established a canonical way of representing this information in a treebank. Noun-incorporation in colonial Nahuatl was quite productive, and in contemporary variants as well, though perhaps to a lesser extent. Tuggy (1987) offers an in-depth overview of the different types of noun-incorporation and the meanings formed by it in the Orizaba variant (`nlv`). We follow the recommendation of Tyers and Mishchenkova (2020) in using UD's enhanced dependencies layer to represent the relationship of the incorporated noun to the verb (Example 35).

(35)

```
              ┌────[advmod]────┐
              ↓                │
     ...   simi   nakas-mahma-tok   nakas
     ...   a lot  ear-stand.up-DUR  ear
           ADV        VERB          NOUN
           2           3            3.1
                        └─[obj]─┘
```
"... always perking up its ears."

6. Automatic Parsing Experiment

In this section we explore how well an automated parser performs when trained on our corpus. We hypothesize that the performance should be quite low, given both the relatively small data volume and the

Metric	Original	Normalized
POS	86.6 ± 1.1	88.9 ± 1.4
UAS	74.4 ± 1.3	77.2 ± 1.7
LAS	65.0 ± 1.4	68.1 ± 2.0

Table 4: Results for part-of-speech tagging (accuracy) and dependency parsing (unlabeled and labeled attachment scores) using UDPipe1, trained on both the original and normalized orthography. Results are the average of 10-fold cross-validation with standard deviation.

high level of dialectal and orthographic variability. We only explore the last of these factors, by evaluating the parser (1) with the original orthography, and (2) using the normalized forms with the INALI orthography.

We use UDPipe 1.2 (Straka et al., 2016) to train an averaged perceptron part-of-speech tagger and neural-network transition-based dependency parser.

The results we obtain are about what should be expected given the volume of data and the amount of internal variation in language varieties and genres. We see a consistent performance improvement when normalizing the orthography, though the results are not significantly different taking into consideration the standard deviations across the 10 folds. We expect to see performance improvements with a more recent UD parser system, such as Udify (Kondratyuk and Straka, 2019) or UDPipe 2.0 (Straka, 2018), which leverage multilingual pretraining. We leave these experiments to future work, but note that such improved performance comes at the cost of resource usage and model size.

7. Discussion

We hope to continue developing this corpus by adding more annotated texts from different genres and regions within the Western Sierra Puebla Nahuatl-speaking region, as well as work with other Nahuatl-speaking communities to develop treebanks for other variants. As mentioned, we are interested in evaluating state-of-the-art UD parsers on our corpus, and exploring cross-lingual parsing of other Nahuatl variants. We hope to leverage this corpus to perform quantitative linguistic analysis of Western Sierra Puebla Nahuatl and contribute to descriptive linguistic work on the language.

We have presented a syntactically annotated corpus of a Nahuatl variant spoken in northern Puebla. We describe important properties of the data, and offer an overview of the annotation decisions made a the level of part-of-speech tags and syntactic constructions. Importantly, this work contributes the first UD treebank for an indigenous Mexican language.

8. Bibliographical References

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

Bello García, S. K., Sánchez Lucero, E., Bonilla Huerta, E., Hernández Hernández, J. C.,

Ramírez Cruz, J. F., and Pedroza Méndez, B. E. (2021). Nahuatl neural machine translation using attention based architectures: A comparative analysis for RNNs and transformers as a mobile application service. In *Mexican International Conference on Artificial Intelligence*, pages 120–139. Springer.

Carochi, H. (2001). *Grammar of the Mexican Language with an explanation of its adverbs (1645)*, volume 89 of *UCLA Latin American Studies*. Stanford University Press, Stanford. Originally published 1645.

de la Cruz Cruz, V. (2010). Las cláusulas relativas en el náhuatl de Teposteco, Chicontepec, Veracruz.

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

Eskander, R., Klavans, J. L., and Muresan, S. (2019). Unsupervised morphological segmentation for low-resource polysynthetic languages. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195.

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

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

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

Gutierrez-Vasques, X., Sierra, G., and Pompa, I. H. (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.

Gutierrez-Vasques, X. (2015). Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 154–160.

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

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.

INALI. (2018). Breviario: Norma ortográfica del idioma náhuatl, méxico. (conforme al avance preliminar de la norma de escritura de la lengua náhuatl a nivel nacional).

Kann, K., Mager, M., Meza-Ruiz, I., and Schütze, H. (2018). Fortification of neural morpho-logical segmentation models for polysynthetic minimal-resource languages. *arXiv preprint arXiv:1804.06024*.

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

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. *arXiv preprint arXiv:1904.02099*.

Langacker, R. W. (1977). *Studies in Uto-Aztecan grammar 1*. Summer Institute of Linguistics and the University of Texas at Arlington, Dallas.

Langacker, R. W. (1979). *Modern Aztec grammatical sketches: Studies in Uto-Aztecan grammar 2*. Summer Institute of Linguistics Publications in Linguistics, 56(2).

Launey, M. and Mackay, C. (2011). *An Introduction to Classical Nahuatl*. Cambridge University Press.

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

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

Lynn, T. and Foster, J. (2016). Universal dependencies for Irish. In *Proceedings of the 2nd Celtic Language Technology Workshop*.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Gonzales, A. R., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Lugo, G. G., Ramos, R., et al. (2021a). Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217.

Mager, M., Oncevay, A., Rios, A., Meza Ruiz, I. V., Palmer, A., Neubig, G., and Kann, K. (2021b). Proceedings of the first workshop on natural language processing for indigenous languages of the Americas. Association for Computational Linguistics.

Mithun, M. (1984). The evolution of noun incorporation. *Language*, 60(4):847–894.

Márquez Hernández, E. and Hernández Juárez, J. (2005a). *In pollito non amo niman otlacat*. Instituto Lingüístico de Verano, México, D.F.

Márquez Hernández, E. and Hernández Juárez, J. (2005b). *In tochtli tecacaquini*. Instituto Lingüístico de Verano, México, D.F.

Márquez Hernández, E. (2001). *¿Tleka kimixmatij in masewalten de otomí?* Instituto Lingüístico de Verano, México, D.F.

Márquez Hernández, E. (2003). *Tlaol cuachicauac*. Instituto Lingüístico de Verano, México, D.F.

Márquez Hernández, E. (2005a). *Ticpinitos uan Ticpintzin*. Instituto Lingüístico de Verano, México, D.F.

Márquez Hernández, E. (2005b). *Tiquitini*. Instituto Lingüístico de Verano, México, D.F.

Márquez Hernández, E. (2005c). *Tlahcahcochini*. Instituto Lingüístico de Verano, México, D.F.

Márquez Pérez, U. (2007). *Tlen ocpanoc in Lencho*. Instituto Lingüístico de Verano, México, D.F.

Müller-Eberstein, M., van der Goot, R., and Plank, B. (2021). How universal is genre in Universal Dependencies? Technical Report arXiv:2112.04971, arXiv.

Naranjo, M. G. and Becker, L. (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*, number 155, pages 91–104. Linköping University Electronic Press.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Olko, J. and Sullivan, J. (2015). Empire, colony, and globalization. a brief history of the Nahuatl language. *Colloquia Humanistica*, pages 181–216, 06.

Pharao Hansen, M. (2015). Dialectal variation in contemporary Nahuatl relative clause formation. AIILS Seminar.

Pugh, R., Tyers, F., and Huerta Mendez, M. (2021). Towards an open source finite-state morphological analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 80–85.

Sapir, E. (1911). The problem of noun incorporation in american languages, Apr.

Sasaki, M. (2014). A dialectological sketch of Ixquihuacan Nahuatl. , 35(TULIP):139–170, sep.

Sasaki, M. (2015). A view from the Sierra : the Highland Puebla area in Nahua dialectology. , 36(TULIP):153–165, sep.

Sasaki, M. (2018). In predecible: Hacer tangible la sintaxis nahua. In *Seminarios de Lenguas Indígenas, UNAM*.

Sasaki, M. (2021). *Configurationality in Ixquihuacan Nahuatl*. Ph.D. thesis, University of Tokyo.

Schroeder, P. and Tuggy, D. H. (2010). The consonantal prefixes of San Miguel Tenango Nahuatl, Zacatlán. *Etnografía del estado de Puebla, zona norte*, pages 112–117.

Schroeder, P. (2014). *Gramática del Náhuatl de San Miguel Tenango, Zacatlán, Puebla*. Summer Institute of Linguistics. [Draft publication].

Schroeder, P. (2015). *Phonology of Nahuatl de San Miguel Tenango, Zacatlán, Puebla*. Summer Institute of Linguistics. [Draft publication].

Straka, M., Hajic, J., and Straková, J. (2016). Udpipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Straka, M. (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.

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

Tuggy, D. (1987). La incorporación de sustantivos en el náhuatl. *SIL Mexico Workpapers*, 8.

Tyers, F. and Henderson, R. (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.

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

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