MesoTree: Annotated Linguistic Resources for Quantitative Comparative Linguistic Analysis and NLP in Mesoamerica

Robert Pugh[1,2] **Francis Tyers**[1,2] **Robert Henderson**[3]
[1]Indiana University [2]Kaltepetlahtol, A.C. [3]University of Arizona
{pughrob, ftyers}@iu.edu rhenderson@arizona.edu

Abstract

One aspect of descriptive and documentary linguistic materials that is becoming increasingly important in the information age is that they be searchable, quantifiable, and comparable. In this paper, we describe an effort to create morphosyntactically-annotated corpora for a number of under-served Mesoamerican (and one Mesoamerica-adjacent) languages using Universal Dependencies. We describe the Mesoamerican linguistic area and languages involved in the project, the training and annotation process, and give a status report on the current state of the corpora. Finally, we describe a comparative syntax experiment and train UD parsing models on the data, demonstrating the usefulness of UD for facilitating quantitative, comparative linguistic research and natural language processing systems.

Keywords: treebank, syntax, Mesoamerica, Mexico, Guatemala, Nahuatl, Huave, Seri, Cuicatec, Uspanteko

1. Introduction

Language documentation can offer a foundation for language revitalization projects (Austin, 2021), and provides linguists with valuable data with which to conduct analyses, in both a monolingual as well as comparative, multilingual settings.

Not only is it critical that language documentation materials be made accessible and available for them to be effective (an area that, fortunately, has been emphasized in most recent documentation projects). Another important factor in evaluating the usefulness of language data from the linguist's perspective is its "searchability"[1], i.e. how easily can one consult the data to find specific examples of phenomena, obtain information about their frequency, and compare with other languages.

With the wide adoption of computational tools for linguistic analysis (Mitkov, 2022; Berez-Kroeker et al., 2023), digital searchability is now more than ever an invaluable attribute of linguistic data. Searchable corpora can also positively impact endangered-language communities by facilitating example identification and the production of valuable lexical resources (Mukherjee, 2006).

Unfortunately, searchability is rarely an area of emphasis in many language documentation projects, and decisions about representation and formatting are often decided on a per-language or per-project basis, making comparative work especially difficult and time-consuming.

In the face of these realities, we propose leveraging the Universal Dependencies framework (Nivre et al., 2020) to support the goals of consistent data formatting and linguistically-motivated searchability and comparison.

Figure 1: Map displaying the Mesoamerican linguistic/cultural area in Mexico and Central America.

In this paper, we describe an active effort to annotate five Mesoamerican languages[2] using Universal Dependencies, pulling from both existing and new texts. In an effort to capture the gamut of morphosyntactic phenomena in these languages, we purposefully set out to annotate three distinct textual genres: dictionary/grammar examples, edited prose, and spontaneous speech.

In the remainder of the paper, we (1) provide background on the Universal Dependencies project, the Mesoamerican *Sprachbünd* (See Figure 1[3]), and the MesoTree project, (2) describe the five lan-

[1]Here, we specifically refer to the ability to search for and compare linguistic phenomena recorded in a language documentation archive. This is not to be confused with the equally important question of "discoverability" of the language documentation material itself, as discussed in, e.g. Yi et al. (2022).

[2]As discussed later, four of these are traditionally considered part of the Mesoamerican linguistic area, and the 5th, Seri, is included as a control language.

[3]Image by El Comandante, via Wikimedia Commons, based on works by AlexCovarrubias and Demis Web Map Server. Distributed under CC BY-SA 3.0.

guages included in our project, (3) highlight the project methodology, including events, outcomes, and challenges, (4) summarize the current state of the corpora and demonstrate their utility via quantitative, comparative analysis.

1.1. Universal Dependencies

The Universal Dependencies (UD) framework is a widely-adopted system for morpho-syntactic annotation by linguists, computer scientists, and others. With a stated goal of facilitating "a linguistic representation that is useful for morphosyntactic research, semantic interpretation, and for practical natural language processing across different human languages", the annotation schema prioritizes "parallelism between similar constructions across different languages" (de Marneffe et al., 2021).

Annotations are a directed, acyclic graph of typed syntactic relations over words in a sentence, with each word including morphological information such as lemma, part-of-speech, and morphological analysis. Notably, these also constitute the levels of analysis typically used in NLP pipelines, making a UD treebank a valuable resource for training automated linguistic analysis systems.

Using an established set of tags and features with cross-linguistic support[4], linguistic analyses become easily comparable across treebanks and languages. Consider the following examples from two Mesoamerican languages, Western Sierra Puebla Nahuatl and Uspanteko (from Sasaki (2021) and Palmer (2009), respectively):

(1) Īwah tiyās n Lwīs ?
ī-wān ti-yā-s n Luis ?
3SG.P-with 2SG.S-go-FUT DEF Luis ?
'Are you going to go with Luis?'

(2) Xinb'ee rik'i wálb
x-in-b'ee r-ik'i w-álb
PERF.A1S-go E3S-with E1S-father.in.law
'I went with my father in law.'

Both examples contain an intransitive motion verb with subject agreement, a relational noun (commonly used in Mesoamerican languages) meaning "with", and a noun specifying the possessor of the relational noun. However, this parallel structure is not clear from the original glossed text. The motivations for the different glossing standards are language specific, and other discrepancies may understandably stem from differences in data source and reflect distinct time periods or grammatical traditions. In Figure 2, showing a UD analysis of the two examples, the linguistic structures within the data become much more easily accessible to comparative analysis.

Due to its thorough yet flexible set of guidelines for annotating text in a way that is both sufficient for language specificity while also cross-linguistically consistent, the UD framework has garnered a great deal of support due to its ease-of-use for quantitative linguistic analyses (Kiss and Thomas, 2019; Tyers and Henderson, 2021), cross-linguistic comparisons (Naranjo and Becker, 2018; Levshina, 2019), and utility for training natural language processing (NLP) pipelines to automatically annotate new input texts. At present there are 339 UD treebanks for 186 languages.

1.2. Mesoamerica as a Linguistic Area

Mesoamerica, a region typically defined as spanning from central Mexico in the North through Central America to the South, has long been recognized as an area of large-scale cultural diffusion (Kirchhoff, 1952). Campbell et al. (1986) systematically investigate the languages spoken in the area and propose its classification as a linguistic area (*Sprachbünd*) given the large number of linguistic features that appear to be shared by diverse, geographically-adjacent languages through language and cultural contact.

Subsequently, a number of linguists have affirmed the validity of Mesoamerica as a linguistic area and shown a great deal of interest in which features define it, which languages should be included or excluded from the area, and the processes of feature diffusion (Munro, 2017; Brown, 2011). Despite ample linguistic work in the area, research in Mesoamerican linguistics unfortunately often relies on disparate resources such as grammars or small, single language corpora. By creating UD treebanks for these languages in areal contact, we can more easily represent features of the Mesoamerican linguistic area that have traditionally been difficult and time-consuming to quantify and compare across disparate grammars.

1.3. The MesoTree Project

This paper describes an ongoing effort to create annotated morphosyntactic corpora, both by annotating existing published texts and by the creation of novel texts for annotation, for four Mesoamerican languages (Cuicatec, Huave, Nahuatl, and Uspanteko) and one Mesoamerica-adjacent language (Seri) in order to better facilitate comparative and quantitative areal linguistics research, natural language processing research on an underrepresented language group, and the development of language technology applications for these languages.

[4]Note, however, that while there is a large set of shared morphological features, language-specific features are also allowed.

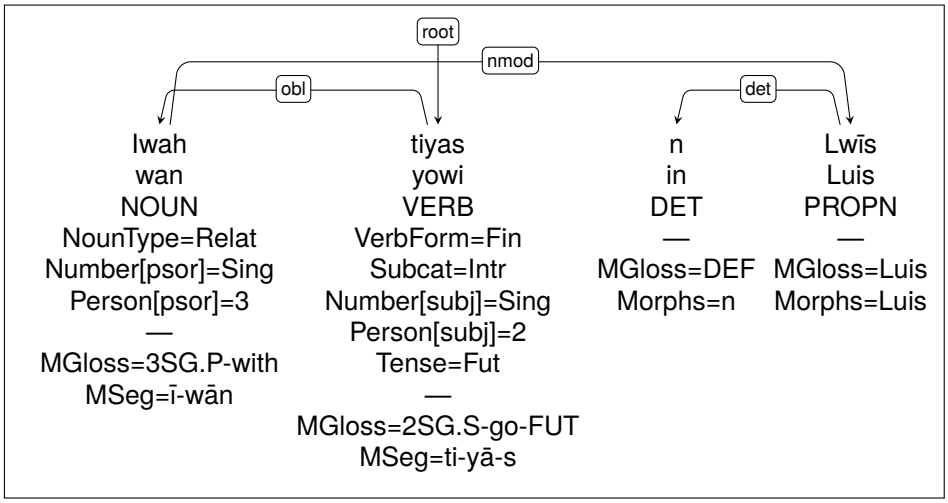

Figure 2: Examples of the UD dependency trees for the sentences in (1) (Top) and (2) (Bottom) using the Universal Dependencies framework. Note that the dependency representations encode not only the same morphological information as the interlinear glosses, but also provide syntactic relationships between the words, facilitating the identification of parallelism between these structures. In both sentences, a possessed noun with a 3rd-person possessor is annotated as an oblique (`obl`) child of the verbal `root`, and a modifying nominal dependent (`nmod`). Note the use of the language-specific morphological feature `NounType=Relat` to capture the unique morphosyntax of Mesoamerican relational nouns.

1.4. UD Treebanks for Languages of the Americas

The UD project includes a large and rapidly-growing set of treebanks in diverse languages, but has until relatively recently consisted of primarily Eurasian languages. Recent years have seen an increase in the representation of indigenous languages of the Americas in UD, including treebanks for Mbyá Guaraní (Thomas, 2019), Nheengatu (de Alencar, 2024), Ika[5] and a number of very small corpora (less than 2,000 tokens) for various languages of Brazil (Akuntsu, Apurina, Guajajara, Kaapor, Karo, Madi, Makurap, Munduruku, Paumari, and Tupinamba) and Paraguay (Guaraní). Karson and Coto-Solano (2024) describe work on a Universal Dependencies-annotated treebank, specifically morphological tagging, for Bribri.Tyers and Henderson (2021) published the first UD treebank for a Mesoamerican language, the Mayan language K'iche', consisting of approximately 10,000 tokens. Since then, treebanks of a similar size for two varieties of Nahuatl have also been added to the UD project: Western Sierra Puebla Nahuatl (Pugh et al., 2022) and Highland Puebla Nahuatl (Pugh and Tyers, 2024). These are currently the only UD treebanks for Mexican languages. The UniMorph project (Batsuren et al., 2022), which aims to provide a cross-linguistically-robust framework for morphological annotation, also includes some

[5] https://github.com/UniversalDependencies/UD_Ika-ChibErgIS

Mesoamerican languages, including varieties of Amuzgo, Chatino (Cruz et al., 2020), Mixtec, and Otomí.

2. Languages

We focus on four genetically unrelated languages spoken in what is commonly considered Mesoamerica: Cuicatec, Huave, Nahuatl, and Uspanteko, and a Mesoamerica-adjacent language, Seri, spoken in Mexico's northern state of Sonora.

2.1. Cuicatec

Cuicatec (*dibaku*, *dbaku*, *dubaku*, ISO 639-3 `cux`) is a highly-endangered Oto-Manguean language spoken in the southern Cañada region of the state of Oaxaca, Mexico, and is the least-studied language in the Mixtecan branch. In this project we focus on Tepeuxila Cuicatec (*dbaku*), spoken in the communities around San Juan Tepeuxila by approximately 8,680 people. There is also substantial internal variation within the variant described as Tepeuxila Cuicatec, a fact that we discuss in Section 3. Relevant existing grammatical work includes the grammar of Belmar (1902), Davis and Walker (1955)'s study of the morphology of the Concepción Pápalo variety, Bradley (1991)'s study of the syntax of Concepción Pápalo, and a more recent study of tonal morphology of the Santa Maria Pápalo variety (Bradley, 1991) and a tonal study of San Juan Tepeuxila variety (San Giacomo Trinidad, 2017). The morphosyntax of Tepeuxila Cuicatec is largely underdocumented.

2.2. Huave

Huave (ISO 639-3 `huv`) is a language isolate consisting of four languages spoken in the Isthmus of Tehuantepec in Oaxaca in the south of Mexico. INALI (2009) defines two principal languages in the grouping: Eastern and Western Huave. In this project we focus on the latter language, specifically that spoken in the community of Huave de San Mateo del Mar (*Ombeayiüts*), which has about 11,000 speakers. Linguistic work on Huave includes a descriptive grammar (Stairs and de Hollenbach, 1981) and a dictionary (Stairs and Stairs, 1981). Some aspects of Huave syntax have been explored, such as the syntax of the noun phrase (Herrera Castro, 2010), person-number marking and valency-changing operations in the verb (Herrera Castro, 2016), but many syntactic phenomena (e.g. clausal subordination) remain undocumented and little understood.

2.3. Nahuatl

Nahuatl is a group of related languages in the Nahuan branch of the Uto-Aztecan language family. Our project is focused on the Western Sierra Puebla variety (ISO 639-3 `nhi`), [6] spoken by approximately 17,100 people in the municipalities of Zacatlán, Ahuacatlán, and Tepetzintla, in Puebla's northern sierra. Compared to other Nahuatl variants, the Western Sierra Puebla variant has received relatively little focus from linguists. The Summer Institute of Linguistics produced an unpublished grammatical sketch and a few dozen children's stories, and Sasaki (2014) provides a minimal grammatical sketch of the communalect spoken in Ixquihuacan, Ahuacatlán, as well as a dissertation about non-configurationality that draws heavily on data collected in this region (Sasaki, 2021). A 30,000-token corpus of spontaneous speech recordings and transcriptions, including word-level language annotations and spontaneous/normalized orthography pairs, was recently published for the variety (Pugh et al., 2025).

2.4. Seri

Seri (*Cmiique iitom*, ISO 639-3 `sei`) is an endangered language isolate (Marlett, 2007), spoken by approximately 900 speakers (Ethnologue 2007 estimate) in the state of Sonora (northwest Mexico), in two villages on the coast of the Gulf of California: El Desemboque (Haxöl Iihom) and Punta Chueca (Socaaix). Seri exhibits complex verbal morphology, displaying a large amount of allomorphy and a high degree of paradigmatic variety (studied in detail in Marlett (2016), a comprehensive grammar, and Baerman (2016)). In addition to the descriptive grammar, a dictionary has also been published. (Moser and Marlett, 2010). There is very little work on syntactic structure of the language beyond Marlett's excellent descriptive grammar.

2.5. Uspanteko

Uspanteko (*Tz'unun Yolooj*, ISO 639-3 `usp`), spoken natively by between 1500-4000 people (Richards and Macario, 2003) in the central highlands of Guatemala, in the city of San Miguel Uspantan and in several surrounding villages, particularly Las Pacayas (Us Maldonado, 2009), belongs to the K'ichean branch of the Mayan language family (Bennett et al., 2016; Aissen et al., 2017).

There is a descriptive grammar of Uspanteko (Can Pixabaj, 2007) and a dictionary (Méndez, 2007). The syntax of Uspanteko is essentially completely unexplored beyond basic description of topic, focus, and question movement.

[6] Alternatively Zacatlán-Ahuacatlán-Tepetzintla Nahuatl

Language	Trees	Tokens
Cuicatec	1,487	7,443
Huave	275	2,260
Nahuatl	2,132	10,466
Uspanteko	2,007	14,676
Seri	367	2,604
Total	6,268	37,449

Table 1: Current corpus volumes by trees and tokens for the five target languages. At this point, three of the five languages have either surpassed or are well on the way to completing the first phase goal of 10,000 tokens. For Nahuatl, annotators have begun annotating fiction texts. For Uspanteko, the collection of fiction texts is ongoing, and annotators have started annotating spontaneous speech transcriptions.

3. Methodology

This section summarizes the training of language expert annotators, the corpus creation process, and some of the challenges faced.

3.1. Text Collection

The project aims to annotate morphosyntactic phenomena across a diverse range of domains and registers. As such, it is broken into three phases, defined by the genre of the texts to be annotated: (1) dictionary/grammar examples, (2) edited prose/fiction, and (3) transcribed spontaneous speech. The work described in the present paper largely corresponds to the first phase. Each phase has a minimum target of 10,000 tokens.

When pre-existing text exists for a language, we use that. For the first phase, we use example sentences from dictionaries, grammars, or other pedagogical publications (Huave (Stairs and Stairs, 1981), Seri (Marlett, 2016), Nahuatl (Brockway, 1984), and Uspanteko (Méndez, 2007)).

Additionally, in some cases (Cuicatec and Nahuatl), at least some of the corpus consists of original example sentences written by language experts. In the case of Cuicatec, the entire corpus is original texts written by Cuicatec-speaking collaborators. The motivation for this decision was the fact that the pre-existing dictionary sentences initially contemplated for the annotation project are written in a different subvariety of Tepeuxila Cuicatec, different than that spoken by the Cuicatec annotators. Similarly, since there is no dictionary for Western Sierra Puebla Nahuatl, and the existing pedagogical material only contains a few hundred sentences, we supplemented this with original example sentences. In both cases, the goal was to produce texts that matched that of the other languages in the project with respect to genre (in this case, dictionary or grammar examples). Therefore, authors were instructed to make a list of words and create an illustrative sentence using each, in effect creating dictionary example sentences.

Additionally, for both Cuicatec and Nahuatl, the project collaborators translated from Spanish a set of approximately 600 example sentences from the Archive of Indigenous Mexican languages (*Archivo de lenguas indígenas de México*, *ALIMG*)[7], typically used for elicitation in documentary linguistics research. It is important to note that, although both Nahuatl annotators are speakers of the Western Sierra Puebla variety, they belong to distinct communities and have distinct subvariants/communalects. For the ALIMG sentences, we asked them each to produce a translation consistent with their community's language variety.

For Nahuatl and Huave, finite-state morphological analyzers have been developed in (Pugh et al., 2021) and (Tyers and Herrera Castro, 2023), respectively. We thus used those, along with some disambiguation rules, to pre-annotate the corpora for those languages with lemma, UPOS, and morphological analysis.

3.2. Training and Annotation

In November, 2023, the project launched with a one-week, in-person course in Mexico City on fundamental concepts in linguistics and Universal Dependency annotation. The course was attended by between one and three language experts from each of the project's target languages. These attendees were primarily native-speaking teachers and/or language activists interested in learning about and contributing to linguistic research for their language. The group also included two non-native-speaker linguistics students at the National Autonomous University of Mexico (UNAM) specializing in one of the five target languages. The content of the course covered fundamentals of morphology and syntax, areal linguistics and the Mesoamerican linguistic area, a deep-dive on the Universal Dependencies project, and a substantial amount of hands-on annotation work. Annotation continued after completion of the course, both asynchronously and in regularly-scheduled video-calls with the project leads. A subsequent 2-day annotation workshop was held in the Summer of 2025 in Mexico City, where the annotators for each language convened for in-person annotation and discussion. All annotation was carried out communally with final annotations decided via group consensus, not as individual efforts. As such, we do not track inter-annotator agreement.

[7] https://cell.colmex.mx/proyecto/archivo-de-lenguas-indigenas-de-mexico/descripcion

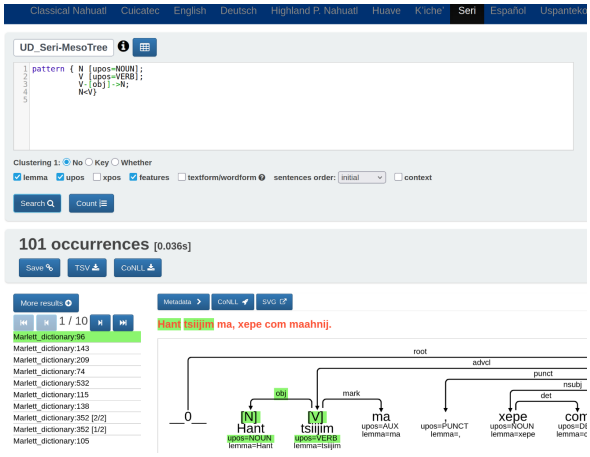

Figure 3: Screenshot of the Grew Match web interface set up with the five target languages, as well as a few others for comparison. This example shows a query for an object noun that precedes a verb.

Annotation is performed via a custom web application[8] that combines the sentence-level annotation interface of UD Annotatrix (Tyers et al., 2017) with the ability to create and manage large-scale, multi-annotator projects offered by Arborator-Grew (Guibon et al., 2020).

3.3. Corpora

The corpus-creation efforts have focused on annotating dictionary/grammar example sentences and, after reaching 10,000 annotated tokens, will expand to increasingly more syntactically-complex sentence types, namely sentences from edited prose and spontaneous speech. Table 1 shows the number of annotated tokens and trees presently. In this case, a "complete" tree refers to, minimally, UPOS and labeled syntactic relations. For some languages, like Nahuatl and Huave, additional morphological information is also included.

The corpora are hosted on a custom instance of the Grew-Match (Guillaume, 2021) web interface, enabling annotators and other researchers to explore syntactic patterns in the corpora via a simple graph query language (See Figure 3). For example, the syntactic pattern highlighted in Figure 2 can be captured with the following query:

3.4. Source and Genre

The first phase of the project has been aimed at annotating dictionary/grammar example sentences, but there were only a small number of such examples for `nhi` and none in the appropriate subvariant of Cuicatec. Therefore, for these two languages we

[8] https://ehecatl.cl.indiana.edu/grew-match

```
pattern {
      RN [ upos=NOUN,
NounType=Relat ] ;
      N [ upos=NOUN ] ;
      V [ upos=VERB ] ;
      V -[obl]-> RN ;
      RN -[nmod]-> N ;
}
```

Figure 4: Grew search pattern for identifying oblique relational nouns with a nominal complement, a pattern shown in Figure 2.

aimed to collect sentences of the “example" genre by (1) having the ALIMG language documentation sentences translated and (2) asking our language consultants to produce new sentences by making a list of words and coming up with an example sentence for each. The motivation was that this process approximates the production of dictionary examples.

In order to evaluate whether these approaches in fact produce texts with similar characteristics as true dictionary example sentences, we compare the different sourcing methods with respect to sentence length and number of clauses per sentence.

The results can be seen in Figure 5. In the left-hand plot we can see that median length varies between five and eight tokens per sentence, and the “true" dictionary/grammar sentences (in green) appear to have a wider distribution of sentence length than the annotator-produced sentences and significantly more so than the ALIMG sentences. It is interesting to note that, though `nhi` and `cux` both translated from the same set of ALIMG sentences, the length for `nhi` sentences appears to be less. This could be due to the fact that Nahuatl can have especially complex morphology, and often a multiword Spanish sentence can be translated with just a single word.

With respect to the number of clauses per sentence, there appears to be much more uniformity across languages and across sources, with most of the sentences from each source having 2 clauses or less. This suggests that, while superficially our sourcing method may produce sentences that are more homogenous with respect to sentence length when compared to actual dictionary examples, they generally match the level of syntactic complexity of the actual dictionary/grammar examples, making the method viable for cross-linguistic comparative syntactic analysis.

3.5. Challenges

For some of the languages, we experienced unexpected challenges that delayed progress. For example, as briefly mentioned in Section 3.1, the

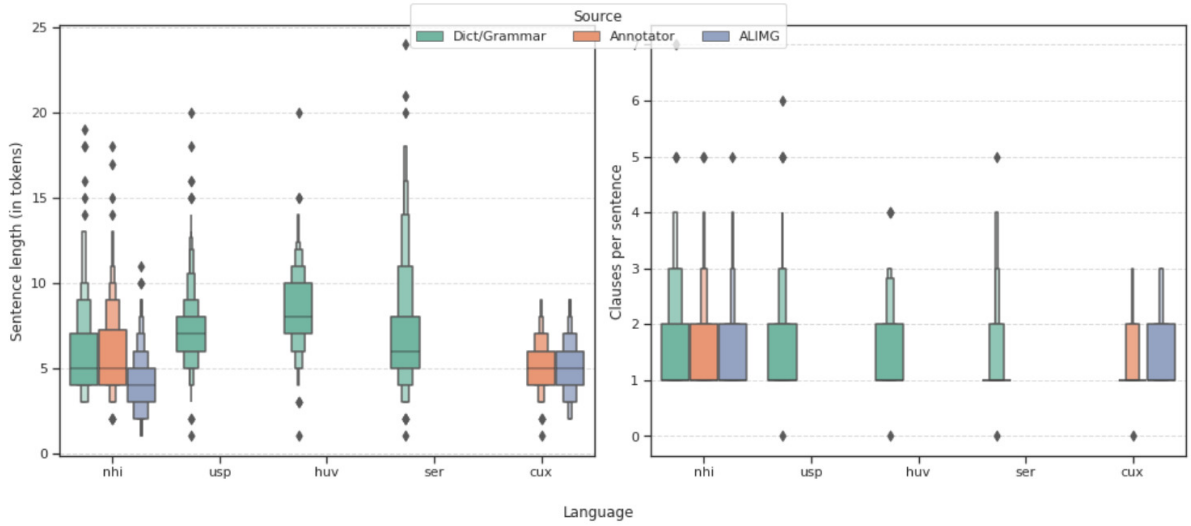

Figure 5: A comparison of the different sources in the "examples" genre. When it was available (for `usp`, `sei`, and `huv`), we used example sentences from a dictionary. In the case of `nhi`, we used sentences from a pedagogical phrase book. For both `nhi` and `cux`, no dictionary examples are available, and we used (1) invented example sentences from the annotators and (2) translations of the ALIMG language documentation sentences.

sentences that were collected initially for Cuicatec ended up being from a different subvariety of Tepeuxila Cuicatec than that of the speakers working on the project. Furthermore, our Cuicatec collaborators were unfamiliar with the orthography used in the dictionary. As a result, it was more time-effective for the speakers to propose sentences in their own variety, written according to their preferred orthographic norms rather than for them to translate the previously collected sentences. We considered automatic conversion of the orthography, but it was complicated by a number of factors, including distinct methods of encoding nasalization and tone.

In general, a major challenge for the project has been unstable access to electricity and/or internet for the annotators living in rural and remote areas of Mexico. Though our annotation platform does not require frequent server-side communication (only to load a treebank and save annotations), it does rely on an internet connection to save progress. Annotation video conferences were frequently ended abruptly due to poor connection or sudden electricity/internet outages.

This issue was particularly prominent for our Huave collaborators in Oaxaca. In an effort to help relieve this problem, we hosted multiple in-person annotation workshops in areas with reliable connection, lasting between one and three days each. In the case Huave, this is where the majority of annotation progress was made.

Progress on Seri annotation has also faced a number of difficulties. Given the small speaker population (900 total, about 300 in the community in which we have contacts), it is challenging to find language experts able to dedicate sufficient time to annotation. Our first annotator is a local bilingual school teacher who unexpectedly had to do national service supporting Seri education. Despite this, she was able to make significant progress upon completing national service.

4. Analysis

In this section we demonstrate two potential use cases for the MesoTree corpora. First, we use the corpora to perform a straightforward quantitative analysis and comparison of syntactic tendencies across (a subset of) Mesoamerican languages, comparing the word-order variability of subjects and objects in our five treebanks along with three pre-existing treebanks of Mesoamerican languages. Next, we examine the utility of the MesoTree treebanks for training NLP systems by training and evaluating multi-task UD parsing models on three of the five languages.

4.1. Word-order Entropy

An important intended use case for the MesoTree treebanks is the facilitation of comparative syntactic analysis. Here, we explore one example of such analysis, using “word-order entropy" (Levshina, 2019). We use Shannon’s Entropy (Equation 1), the average amount of information required to describe the events of a random variable, to quantify the word-order variability of clauses with overt subjects and objects.

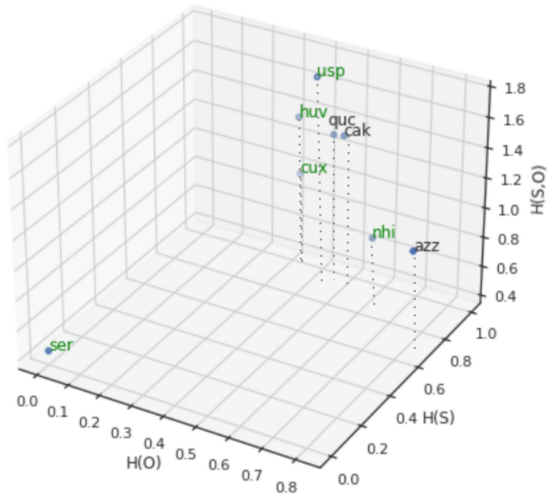

Figure 6: A plot of Mesoamerican languages with annotated corpora as a function of the entropy of nominal subject and object order. Languages whose datasets were created as a part of the presently-described project are labeled in green.

$$H(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x) \quad (1)$$

Specifically, we measure the entropy of (1) verb-subject order, (2) verb-object order, and (3) verb-subject-object order. In this case, higher entropy indicates greater syntactic unpredictability.

For thoroughness, we include three previously-annotated Mesoamerican treebanks (see Section 1.4) in the analysis. Specifically, in addition to the five target-language corpora described in Section 3.3, we include corpora for K'iche' (Tyers and Henderson, 2021) and Highland Puebla Nahuatl (Pugh and Tyers, 2024), each totaling approximately 10,000 tokens, as well as a small, unpublished corpus of Kaqchikel (approximately 1,000 tokens) created by one of the authors. We combine the existing Western Sierra Puebla Nahuatl treebank (Pugh et al., 2022) with our project's annotations for the same language.

The results are listed in Appendix A and visualized in Figure 6. There is an apparent pattern among the Mesoamerican languages, with most having verb-subject entropy near 1.0 (maximally unpredictable, or effectively free order), and slightly more predictable orders with explicit objects. There is much more variance along the Subject-Object order entropy. The Mesoamerican-adjacent control language, Seri, has much more consistent word orders and is a clear outlier.

It is also note-worthy that the three Mayan languages, Uspanteko (usp), K'iche' (quc), and Kaqchikel (cak), and the Nahuatl languages Western Sierra Puebla Nahuatl (nhi) and Highland Puebla Nahuatl (azz) seem to form groupings. The fact that the Nahuatl languages appear slightly separated from the remaining Mesoamerican languages could be due to genre: The azz treebank and the older nhi treebank (which was combined with the MesoTree treebank) contain both edited prose and a large number of spontaneous speech transcriptions. Once the treebanks for all three genres are completed, we plan to re-run this analysis to analyze both linguistic and genre-based factors.

4.2. Training Automatic UD-Parsing Models

Lang	Lemma	UPOS	UAS	LAS
usp	0.94	0.94	0.81	0.75
nhi	0.84	0.94	0.86	0.81
cux	-	0.90	0.86	0.76

Table 2: Comparison of Lemmatization, UPOS tagging, and dependency parsing for the Uspanteko (usp), Western Sierra Puebla Nahuatl (nhi), and Tepeuxila Cuicatec datasets. Since the morphological analyzer for Cuicatec is still under development, the lemma annotation is incomplete, and we omit lemma accuracy.

Finally, given our stated interest in leveraging the MesoTree corpora for building NLP systems for Mesoamerican languages, we train and evaluate UD parsing models the three languages with the most annotated data, usp, nhi, and cux.

We use MaChAmp (van der Goot et al., 2021) to fine-tune multilingual BERT (mBERT) embeddings[9] on each UD task. 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 the experiment, we split each treebank randomly into a train (85% of the sentences) and eval (15% of the sentences) set.

5. Concluding Remarks

The MesoTree project aims to produce high-quality linguistically-annotated corpora that enable and encourage computer-aided linguistic analyses and language technology development for Mesoamerican languages. Though work on the project is still underway, the value of producing consistently-annotated corpora across multiple unrelated lan-

[9]We use the bert-base-multilingual-cased model.

guages in areal contact is already apparent, as demonstrated by the quantitative comparative word-order analysis and the training of automatic parsing models described in this paper.

Future work will focus on expanding these corpora to include more complex registers, such as edited prose and spontaneous speech, ultimately providing a foundation for language technology that serves both the global research community and the speaker communities.

Acknowledgements

We would like dedicate this work to our dear colleague, Debora Perales Morales, a linguist and speaker of the Seri language, who passed away unexpectedly this fall from sudden illness. She was responsible for the Seri annotations described in this paper, but what a small thing compared to her tireless advocacy for her language and her community. It is hard to know how to move forward having lost such a great mind. We must find a way in her honor.

We also must recognize NSF grant #2319246/2319247 *Collaborative Research: Syntactically-annotated corpora for endangered languages in areal contact* for supporting this work.

6. Bibliographical References

Judith Aissen, Nora C England, and Roberto Zavala Maldonado. 2017. *The Mayan languages*. Taylor & Francis.

Peter K. Austin. 2021. *Language Documentation and Language Revitalization*, page 199–219. Cambridge University Press.

Matthew Baerman. 2016. Seri verb classes: Morphosyntactic motivation and morphological autonomy. *Language*, 92(4):792–823.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, et al. 2022. UniMorph 4.0: universal morphology. *arXiv preprint arXiv:2205.03608*.

Francisco Belmar. 1902. El Cuicateco-Primera Edición.

Ryan Bennett, Jessica Coon, and Robert Henderson. 2016. Introduction to Mayan linguistics. *Language and Linguistics Compass*, 10(10):455–468.

Andrea L. Berez-Kroeker, Shirley Gabber, and Aliya Slayton. 2023. Recent advances in technologies for resource creation and mobilization in language documentation. *Annual Review of Linguistics*, 9(1):195–214.

David P Bradley. 1991. A preliminary syntactic sketch of Concepción Pápalo Cuicatec. *Studies in the syntax of Mixtecan languages*, 3:409–506.

Earl Brockway, editor. 1984. *Frases en mejicano y español: En el idioma náhuatl de San Miguel de Tenango, Puebla y en español*, primera edición edition. Instituto Lingüístico de Verano, A.C., Mexico.

Cecil Brown. 2011. The Role of Nahuatl in the Formation of Mesoamerica as a Linguistic Area. *Language Dynamics and Change*, 1:171–204.

Lyle Campbell, Terrence Kaufman, and Thomas C Smith-Stark. 1986. Meso-america as a linguistic area. *Language*, pages 530–570.

Telma Angelina Can Pixabaj. 2007. *Gramática uspanteka*. Cholsamaj Fundacion.

Hilaria Cruz, Antonios Anastasopoulos, and Gregory Stump. 2020. A resource for studying chatino verbal morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2827–2831, Marseille, France. European Language Resources Association.

Marjorie Davis and Margaret Walker. 1955. Cuicateco: Morphemics and morphophonemics. *International Journal of American Linguistics*, 21(1):46–51.

Leonel Figueiredo de Alencar. 2024. A Universal Dependencies treebank for nheengatu. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 37–54, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

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.

Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175.

Samuel Herrera Castro. 2010. Alineamiento y frase verbal en Huave de San Mateo Del Mar, Oaxaca. Master's thesis, Escuela Nacional de Antropología e Historia, México, D.F.

Samuel Herrera Castro. 2016. *Sintaxis y semántica de la frase nominal en Huave de San Mateo Del Mar, Oaxaca*. Ph.D. thesis, Colegio de México, México, D.F.

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.

Jessica Karson and Rolando Coto-Solano. 2024. Morphological tagging in Bribri using Universal Dependency features. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 56–66, Mexico City, Mexico. Association for Computational Linguistics.

Paul Kirchhoff. 1952. Mesoamerica: its geographic limits, ethnic composition and cultural characteristics. *Heritage of conquest*, pages 17–30.

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.

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

Stephen A. Marlett. 2007. Las relaciones entre las lenguas "hokanas" en México: ¿cuál es la evidencia? *Memorias del III Coloquio Internacional de Lingüística Mauricio Swadesh*, pages 165–192.

Stephen A Marlett. 2016. Cmiique Iitom: the Seri language. *Unpublished grammar (2016 draft)*.

Miguel Angel Vicente Méndez. 2007. *Cholaj Tzijb'al li Uspanteko: Diccionario bilingüe uspanteko-español*. Cholsamaj.

Ruslan Mitkov. 2022. *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Mary B. Moser and Stephen A. Marlett. 2010. *Comcaac quih yaza quih hant ihiip hac: Diccionario seri-español-inglés*, 2nd edition. Plaza y Valdes editores and Universidad de Sonora.

Joybrato Mukherjee. 2006. Corpus linguistics and language pedagogy: The state of the art–and beyond. *Corpus technology and language pedagogy: New resources, new tools, new methods*, pages 5–24.

Pamela Munro. 2017. The Mesoamerican linguistic area revisited. *Language contact and change in Mesoamerica and beyond*, pages 335–351.

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.

Alexis Mary Palmer. 2009. Semi-automated annotation and active learning for language documentation.

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.

Robert Pugh, Francis Tyers, and Marivel Huerta Mendez. 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.

Robert Pugh and Francis M. Tyers. 2024. 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. 2025. 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.

Michael Richards and Narciso Cojtí Macario. 2003. *Atlas lingüístico de Guatemala*. Editorial Serviprensa Guatemala City.

Marcela San Giacomo Trinidad. 2017. The phonetics and phonology of tone in San Juan Tepeuxila Cuicatec. *Cuadernos de Lingüística de El Colegio de México*, 4(2):83–136.

Mitsuya Sasaki. 2014. A dialectological sketch of Ixquihuacan Nahuatl. *Tokyo University Linguistic Papers*, 35:e139–170.

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

Emily F. Stairs and Elena E. de Hollenbach. 1981. Gramática huave. In Glenn A. Stairs Kreger and Emily F. Scharfe de Stairs, editors, *Diccionario huave de San Mateo del Mar*, pages 283–391. Instituto Lingüístico de Verano, México, D. F.

Glen Stairs and Emily Stairs. 1981. *Diccionario huave de San Mateo del Mar*. Instituto Lingüístico de Verano, México, D.F.

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.

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, 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.

Francis M. Tyers and Samuel Herrera Castro. 2023. Towards a finite-state morphological analyser for San Mateo Huave. In *Proceedings of the 6th Workshop on Computational Methods for Endangered Languages*. [to appear].

Juan Antonio Us Maldonado. 2009. *Monografía Uspanteka*. Comunidad Lingüística Uspanteka, San Miguel Uspantán.

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.

Irene Yi, Amelia Lake, Juhyae Kim, Kassandra Haakman, Jeremiah Jewell, Sarah Babinski, and Claire Bowern. 2022. Accessibility, discoverability, and functionality: An audit of and recommendations for digital language archives. *Journal of Open Humanities Data*.

A. Shannon's Entropy

The entropy calculations for the 8 treebanks (seven for Mesoamerican languages and one Mesoamerica-adjacent) can be seen in Tabl 3.

Language	H(O)	H(S)	H(S,O)
azz	0.81	0.68	1.01
cak	0.48	0.95	1.35
cux	0.32	1.00	0.96
huv	0.32	1.00	1.35
nhi	0.61	0.89	0.79
quc	0.44	0.98	1.31
sei	0.0	0.0	0.44
usp	0.42	0.90	1.74

Table 3: Entropy calculations for the 8 treebanks (seven for Mesoamerican languages and one Mesoamerica-adjacent), corresponding to the visualization in Figure 6