Universal Dependencies for Western Sierra Puebla Nahuatl

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

† Indiana University

∗ Independent

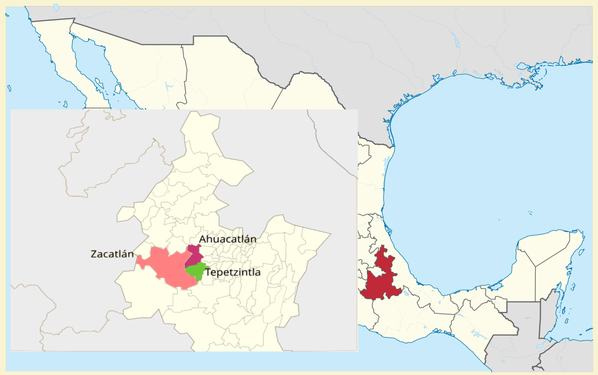
Overview/Contributions

- Syntactically-annotated corpus of an understudied variant of Nahuatl
- Explore UD guidelines with respect to typological and sociolinguistic diversity.
- Contribute to the representation of Mesoamerican languages in UD.

Universal Dependencies

- UD: Consistent schema for morph. and synt. annotation for as many languages as possible.
- Despite great progress, most existing UD treebanks are for Eurasian languages and on edited/published texts.
- Treebank includes all of the major components of a typical NLP pipeline: Tokenization, lemmatization, POS tagging, morphological and synstactic analysis.

Western Sierra Puebla Nahuatl



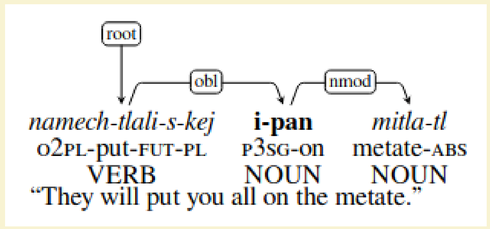
- Nahuatl is a **polysynthetic**, agglutinating Uto-Aztecan language continuum spoken throughout Mexico and Mesoamerica
- Western Sierra Puebla Nahuatl is one of 30+ Nahuatl variants, spoken primarily in the municipalities of Ahuacatlán, Zacatlán, and Tepetzintla
- Approx. 17k speakers

Corpus

- 11 sources, 939 trees, 10,356 tokens.
- All three municipalities represented (Ahuacatlán, Zacatlán, and Tepetzintla)
- Variant-internal linguistic (lexical and morphological) and orthographic variation.
- We include the original form **and** the orthographically-normalized form.

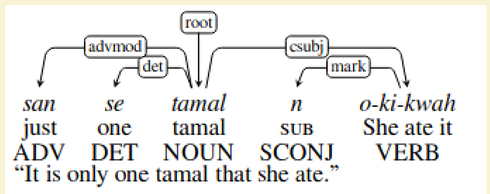
Syntactic Constructions

- **Possession and Relational Nouns**

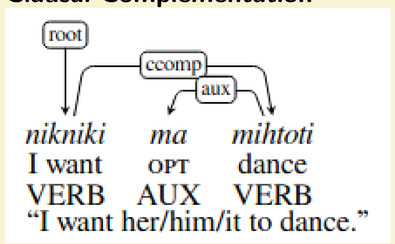


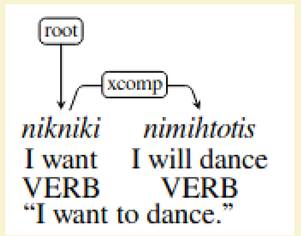
i-tikak-uan *no-papa*
P3SG-shoe-PL P1SG-father
"My father's shoes."

- **Clefting/Focalization**



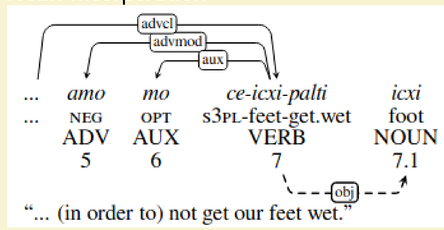
- **Clausal Complementation**





Syntactic Constructions cont.

- Noun Incorporation



Automatic Parsing Experiment

- UDPipe 1.0
- 10-fold cross-validation, POS, UAS, and LAS
- Tested both original and normalized orthography
- **Results**

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: 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.

Future Work

- Explore state-of-the-art dependency parsers on this dataset, including evaluating the effect of different cross-/multilingual training methods.
- Work on similar corpora for other Nahuatl variants, and performing cross-dialectal quantitative analyses.

pughrob@iu.edu