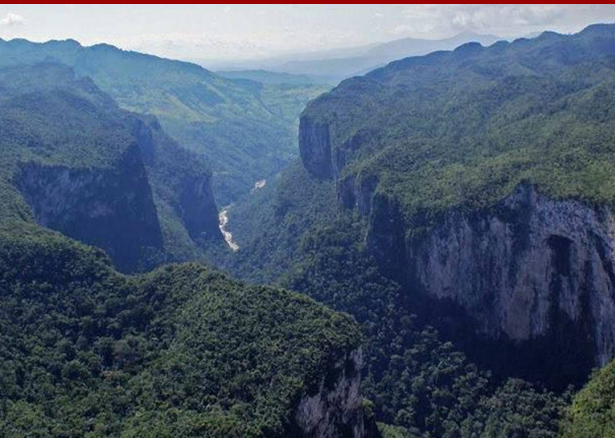


A Universal Dependencies Treebank for Highland Puebla Nahuatl

Robert Pugh & Francis Tyers

Indiana University, Department of Linguistics



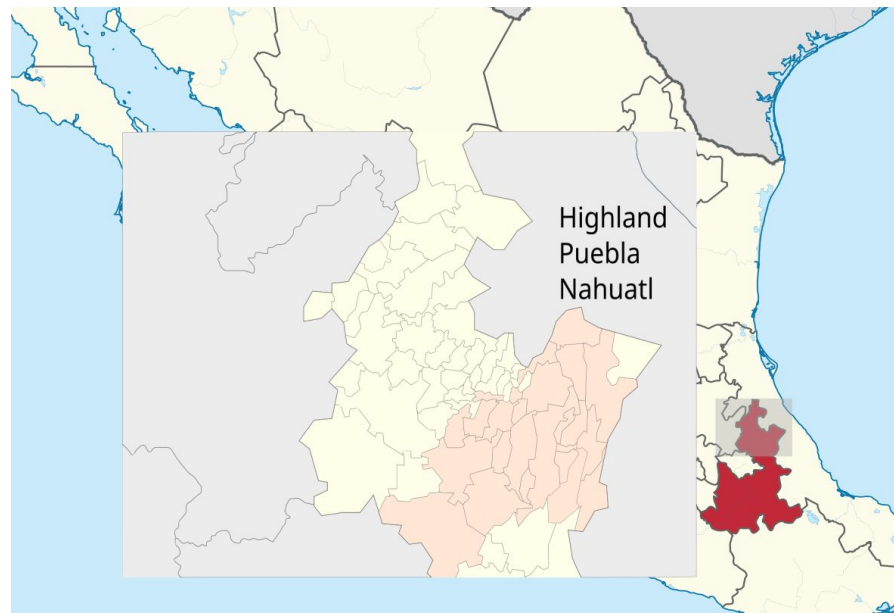


- Highland Puebla Nahuatl (HPN)
- Universal Dependencies (UD) treebanks
- Previous and related work
- Data
- Annotation process and some constructions
- Corpus comparison
- Training a UD parser for HPN.

Highland Puebla Nahuatl



- Nahuatl: **polysynthetic, agglutinating** Uto-Aztec language continuum spoken in **Mexico and Mesoamerica**
- Highland Puebla Nahuatl (azz): 1 of ~30 variants, spoken in the **northeastern Sierra region**, by about **70k speakers in 24 municipalities**
- Rapid language shift to Spanish in most communities



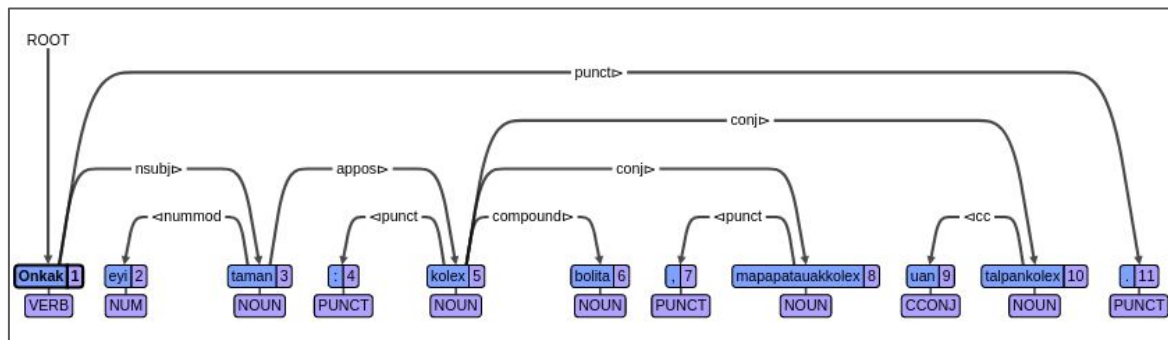
Universal Dependencies



- Framework for **cross-linguistically-consistent morphosyntactic annotation**.
 - useful for NLP applications, linguistic typology research, psycholinguistic research, etc.



161 languages, 283 treebanks



Universal Dependencies



Universal Dependencies





OpenSLR

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

Annotation Process



- Lemmas, UPOS, and morphological features from FST

[morphological analyzer](#)

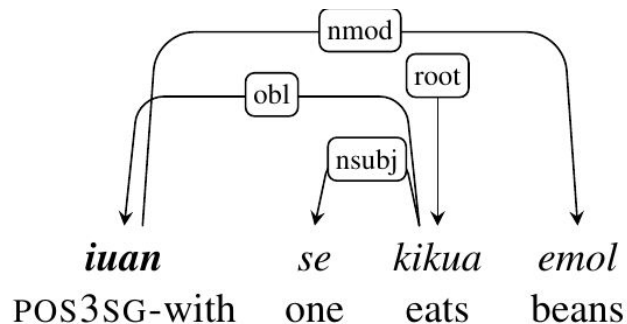
```
^Ixua/<s_sg3>ixua<v><iv><pres>$  
^uan/huan<cnjcoo>$  
^moskaltia/<s_sg3>moskaltia<v><iv><pres>$  
^,/,<cm>$  
^ijuak/ijhuak<cnjsub>$  
^motamiti/<s_sg3>motami<v><iv><and>$  
^peua/<s_sg3>pehua<v><iv><pres>$  
^xochiyoua/<s_sg3>xochiyohua<v><iv><pres>$  
^./.<sent>$
```

- Syntactic heads and relation labels **manually annotated** and then **manually reviewed** by a different annotator, with **discussion**

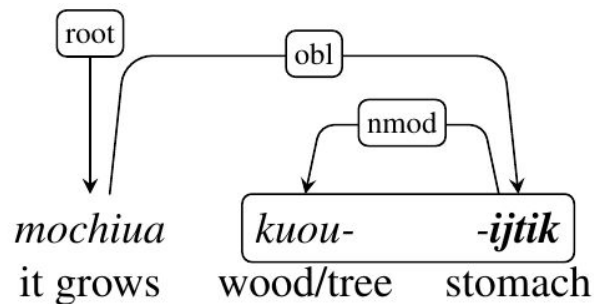
Syntactic constructions



- Relational Nouns
 - possessed or compounded
 - can be disjoint



“It is eaten with beans.”

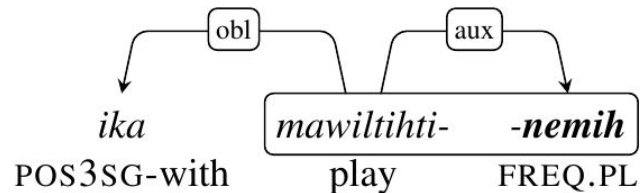


“It grows in the woods.”

Syntactic constructions

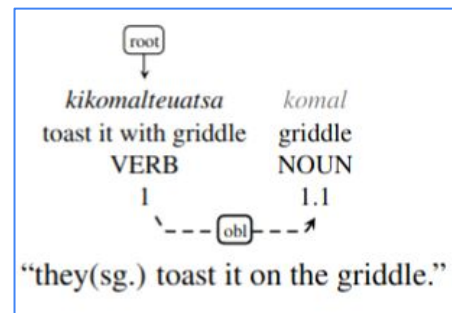
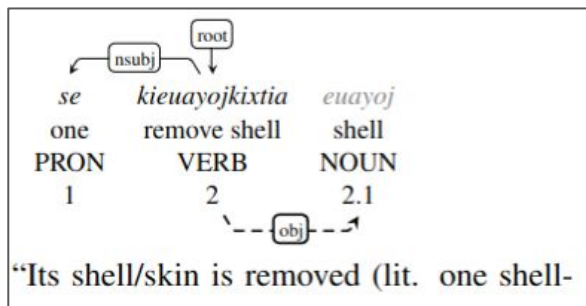


- Verb-Aux compounds



“Children go around playing with it.”

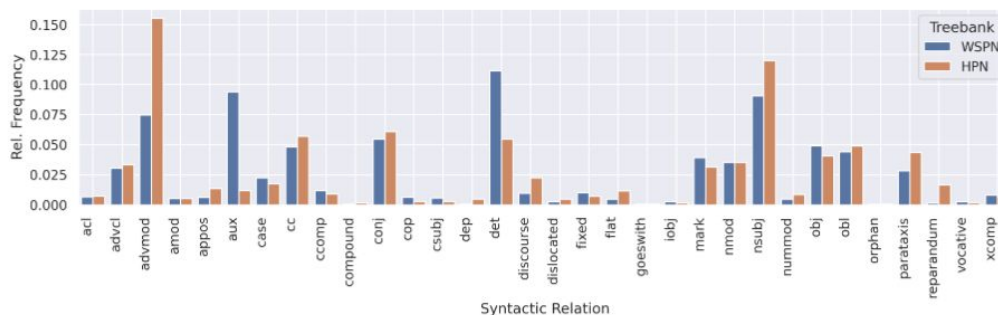
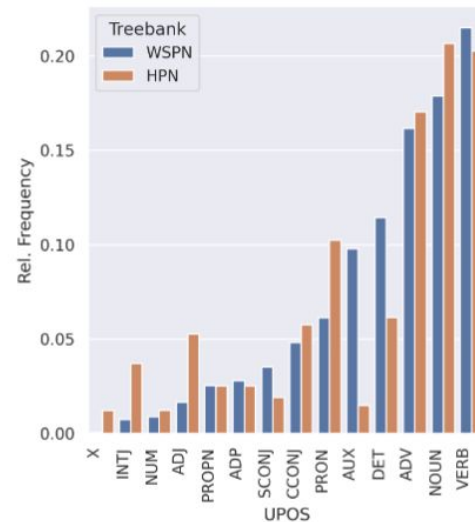
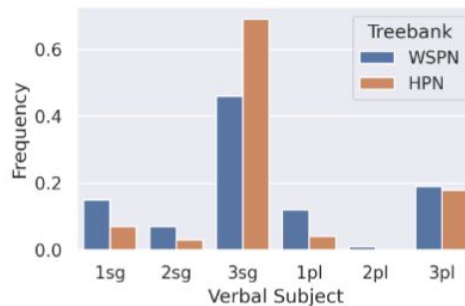
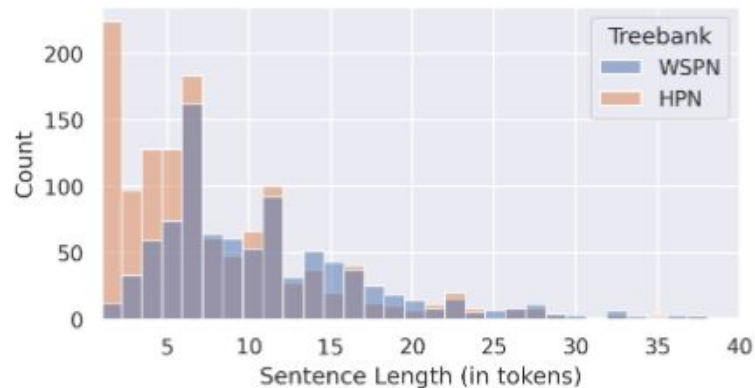
- Noun incorporation (and/or Noun/Verb compounding)



Comparison with WSPN treebank



Comparison with WSPN treebank



HPN Parser

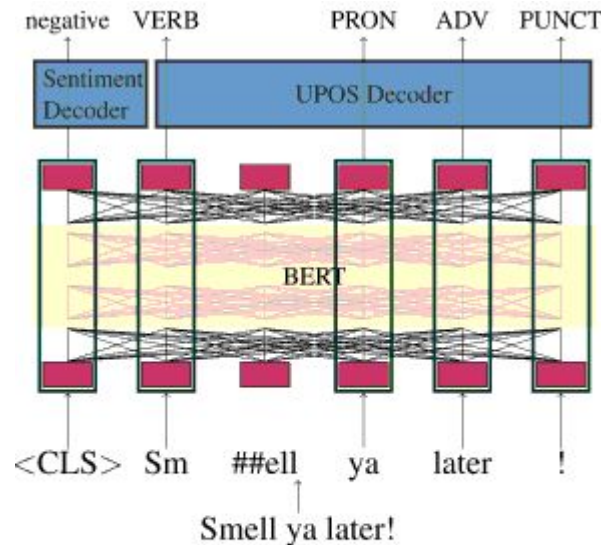


Figure 1: Overview of MACHAMP, when training jointly for sentiment analysis and POS tagging. A shared encoding representation and task-specific decoders are exploited to accomplish both tasks.

<https://github.com/machamp-nlp/machamp>

HPN Parser: Experiments



- How well can a neural parser learn to annotate these texts?
 - 10-fold **cross-validation** on treebank
 - **Accuracy**: lemmas, UPOS, morph features, unlabeled and labeled attachment
- Evaluate **domain-specificity** of parser trained on this data.
 - **Zero-shot prediction** on sample from **unseen domain** (narratives about cultural practices)
 - **Manually post-edit predictions** to get the correct annotations.
 - Omit lemmas and morph (time-consuming to post-edit)

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



- Expand the domains of azz UD treebank
- Continue larger project of creating annotated corpora for Nahuatl varieties
- Quantitative, corpus-based analysis of syntactic dimension of Nahuatl dialectology.

Tasojkamatik

