Experiments in Multi-variant NLP for Nahuatl

Robert Pugh & Francis Tyers

Indiana University, Department of Linguistics

Goals & Research questions

Goal: Automatic UD parsing for as many Nahuatl varieties as possible

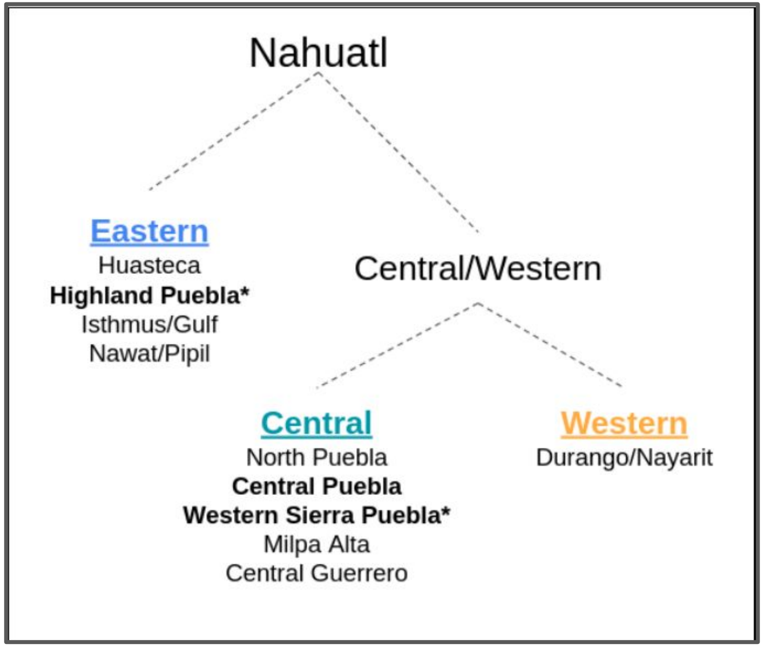
Some questions:

- To what extent do differences between Nahuatl variants effect parsing performance?
- How well do cross-lingual and jointly-trained, multivariant models perform on the task?
- Is it necessary to train a separate model for each variant, or is a multivariant model feasible?
- Can we improve the parser with an auxiliary task using multivariate text?

The Nahuatl language "complex"

- Polysynthetic, agglutinating Uto-Aztecan "language complex".
- ~30 variants spoken throughout Mexico (no standard variety)
- Also a historical, literary "variety" Classical Nahuatl.
- Textual resources:
  - Historical documents, books for some contemporary variants (Axolotl corpus),
  - **UD treebanks (Western Sierra Puebla Nahuatl [nhi], Highland Puebla Nahuatl [azz])**

The Nahuatl language "complex"



The Nahuatl language "complex"

Variants can differ on any level of linguistic analysis:

- **Lexical**: *totoltetl* vs. *tecsistli* "egg", *yetoc* vs. *cah* "to be (cop.)"
- **Phonological**: *e→i*, /tl/ vs /t/ vs. /l/, word-final consonant clusters
- **Morphological**: presence/absence of antecessive o-, suffix -qui in past
- **Syntax**: Relative clauses, VERB-ADV order preferences

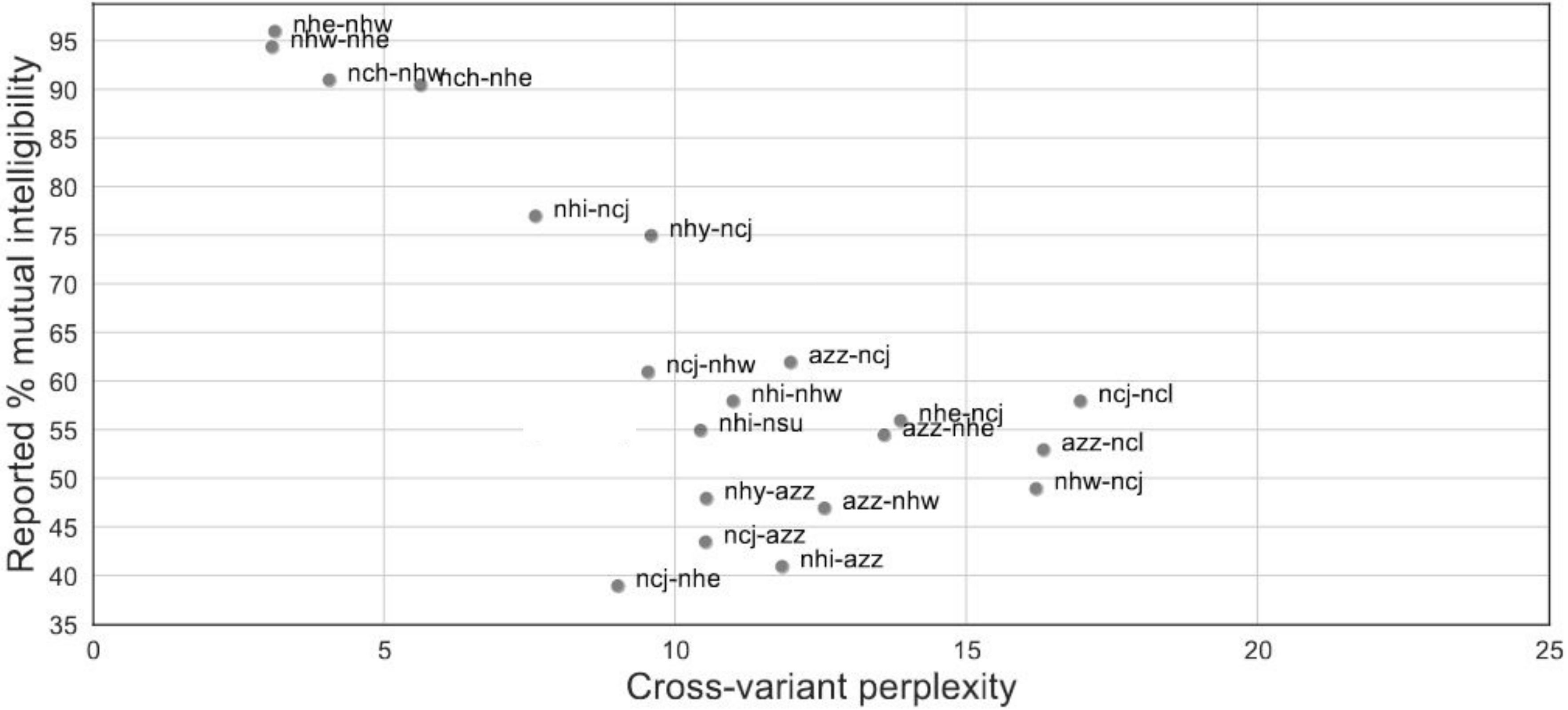
Orthography & Contact phenomena: Highly variable both at the community and individual level.

The Nahuatl language "complex": nhi & azz

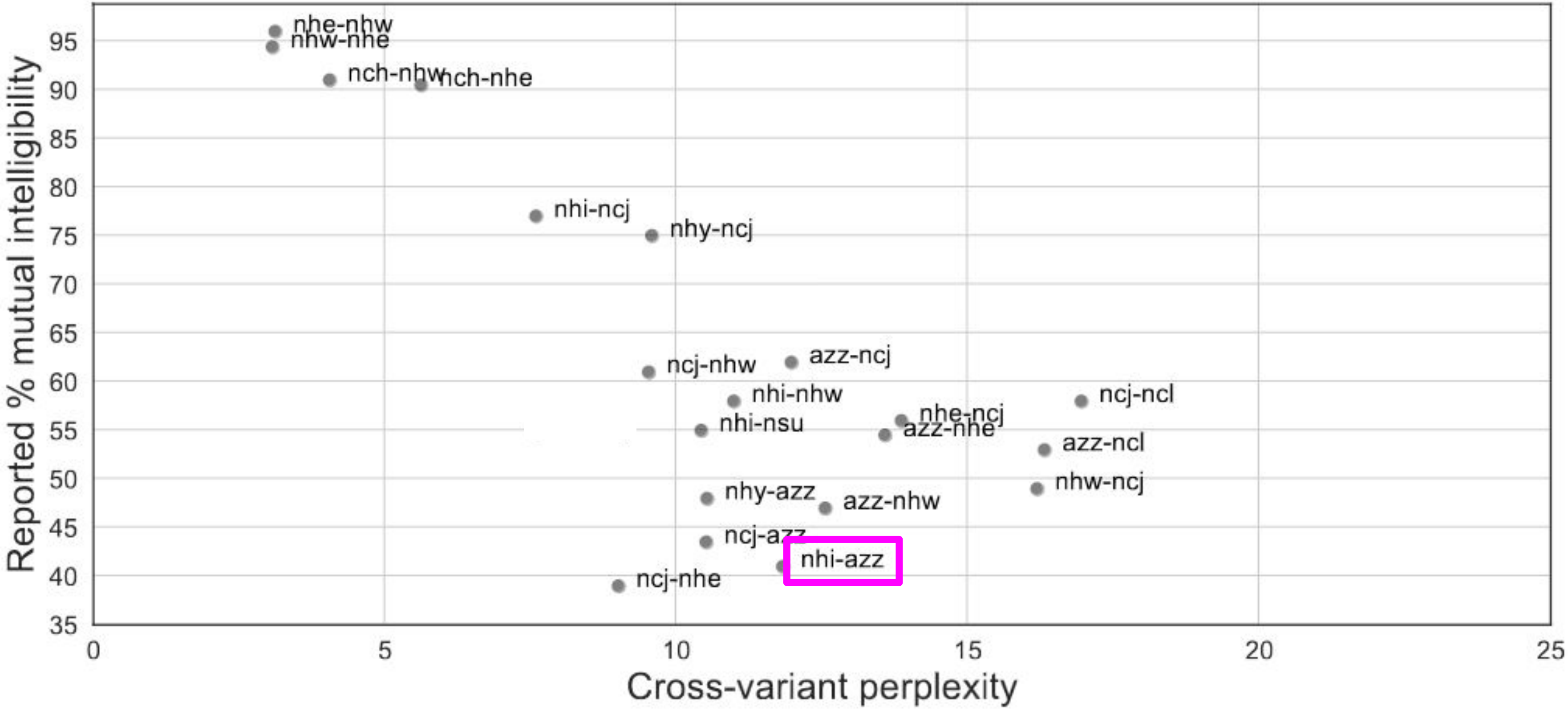
azz	nhi	en
Tepos **teyin** tepaleuia mah **ica** se quita **teyin** amo ueli se quita ica se ixtololo.	Tipos**tl** **tlen** tepaleuia **ica** mo sequita **tlen** amo uili sequita ica se ixtololo.	"Instrument that helps people see what cannot be seen with an eye."
Ocs**e**pa ti**qui**yolitijkej.	Ocs**i**pa oti**c**yolitihkeh.	"We started it up again."

- e vs i
- tepos vs tipostl, teyin vs tlen, ueli vs. uili, ocsepa vs. ocsipa
- syntax of auxiliary "mah/mo" and RelN "ica"
- phonological realization of 3rd-sg object pref /k/

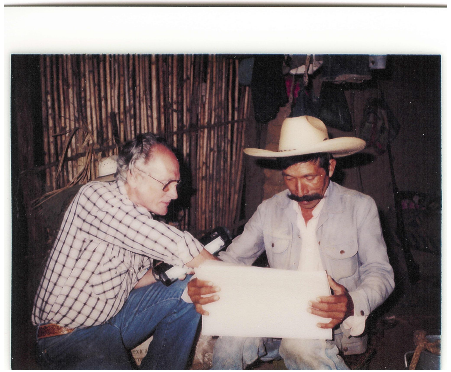
Mutual intelligibility between variants



Mutual intelligibility between variants



Data



Use	Source	Variant	Annotation	Tokens	Sents
train/eval	azz treebank	azz	UD	10,088	1,260
train/eval	nhi treebank	nhi	UD	10,132	909
train only	Axolotl	azz, nci, nhm, nhn, nhw	unlabeled	182,174	13,519
eval only	Casanova stories	ncx	UD	2,355	200

Experiments

- Monolingual
- Cross-variant
- Joint I
- Joint I w/ MLM auxiliary task (Axolotl corpus)
- Joint II (½ volume from each variant)

Experiments

- Monolingual
- Cross-variant
- Joint I
- **Joint I w/ MLM auxiliary task (Axolotl corpus)**
  13k sentences (azz, nci, nhm, nhn, nhw)
- Joint II (½ volume from each variant)

Experiments: model architecture

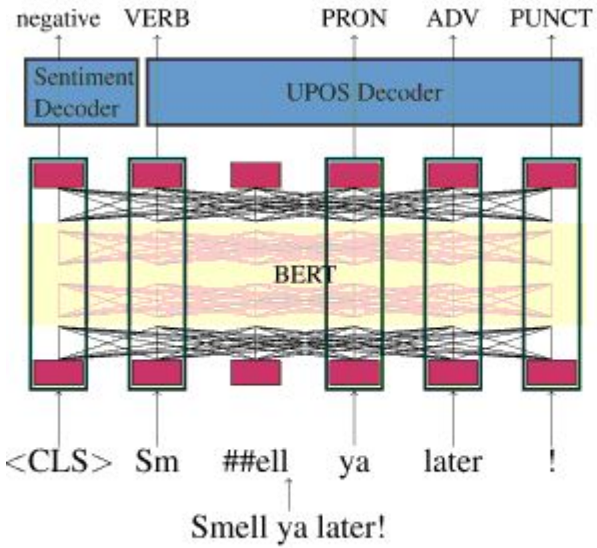
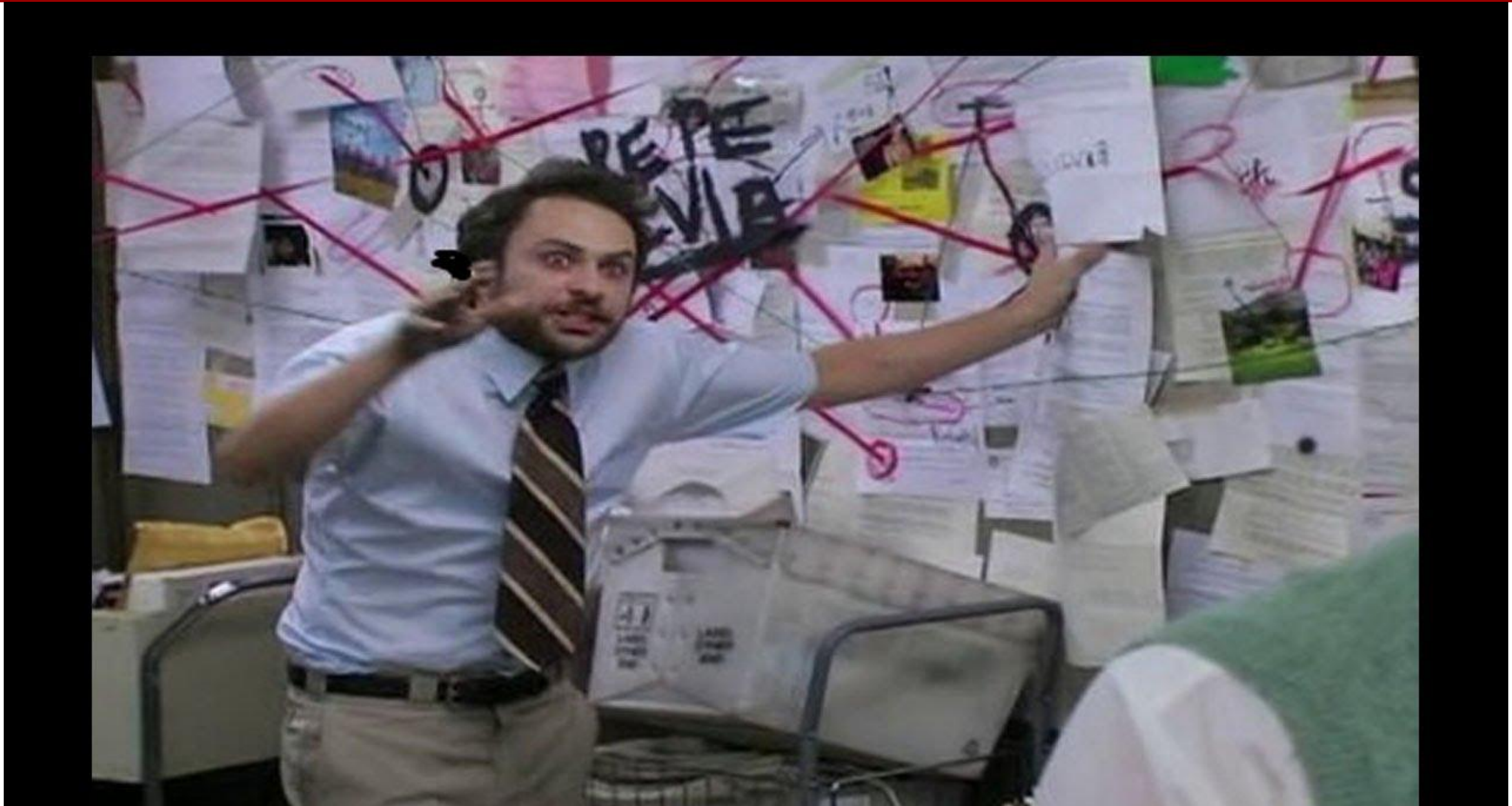


Figure 1: Overview of MACHAMP, when training jointly for sentiment analysis and POS tagging. A shared encoding representation and task-specific decoders are exploited to accomplish both tasks.

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

Experiments: Results & Analysis



Experiments: Results & Analysis

Var.	Experiment	N	Lemma	UPOS	Morph.	UAS	LAS
azz	Mono	1,134	0.92 ± 0.02	0.94 ± 0.01	0.85 ± 0.02	0.84 ± 0.02	0.77 ± 0.03
	Cross	818	0.68 ± 0.02	0.68 ± 0.02	0.39 ± 0.01	0.67 ± 0.02	0.47 ± 0.03
	Joint Adj.	976	0.89 ± 0.01	0.93 ± 0.01	0.75 ± 0.02	0.81 ± 0.02	0.73 ± 0.02
	Joint	1,952	0.92 ± 0.01	0.95 ± 0.01	0.82 ± 0.03	0.85 ± 0.02	0.77 ± 0.02
	Joint+MLM	1,952	0.92 ± 0.01	0.95 ± 0.01	0.82 ± 0.02	0.85 ± 0.02	0.78 ± 0.02
nhi	Mono	818	0.82 ± 0.02	0.93 ± 0.01	0.67 ± 0.02	0.83 ± 0.02	0.74 ± 0.02
	Cross	1,143	0.65 ± 0.02	0.65 ± 0.02	0.44 ± 0.01	0.64 ± 0.02	0.42 ± 0.01
	Joint Adj.	976	0.79 ± 0.02	0.91 ± 0.02	0.60 ± 0.02	0.81 ± 0.02	0.71 ± 0.02
	Joint	1,952	0.82 ± 0.02	0.93 ± 0.01	0.67 ± 0.02	0.84 ± 0.02	0.76 ± 0.02
	Joint+MLM	1,952	0.82 ± 0.02	0.93 ± 0.01	0.68 ± 0.01	0.85 ± 0.02	0.76 ± 0.03

Experiments: Results & Analysis

Var.	Experiment	N	Lemma	UPOS	Morph.	UAS	LAS
azz	Mono	1,134	0.92 ± 0.02	0.94 ± 0.01	0.85 ± 0.02	0.84 ± 0.02	0.77 ± 0.03
	Cross	818	0.68 ± 0.02	0.68 ± 0.02	0.39 ± 0.01	0.67 ± 0.02	0.47 ± 0.03
	Joint Adj.	976	0.89 ± 0.01	0.93 ± 0.01	0.75 ± 0.02	0.81 ± 0.02	0.73 ± 0.02
	Joint	1,952	0.92 ± 0.01	0.95 ± 0.01	0.82 ± 0.03	0.85 ± 0.02	0.77 ± 0.02
	Joint+MLM	1,952	0.92 ± 0.01	0.95 ± 0.01	0.82 ± 0.02	0.85 ± 0.02	0.78 ± 0.02
nhi	Mono	818	0.82 ± 0.02	0.93 ± 0.01	0.67 ± 0.02	0.83 ± 0.02	0.74 ± 0.02
	Cross	1,143	0.65 ± 0.02	0.65 ± 0.02	0.44 ± 0.01	0.64 ± 0.02	0.42 ± 0.01
	Joint Adj.	976	0.79 ± 0.02	0.91 ± 0.02	0.60 ± 0.02	0.81 ± 0.02	0.71 ± 0.02
	Joint	1,952	0.82 ± 0.02	0.93 ± 0.01	0.67 ± 0.02	0.84 ± 0.02	0.76 ± 0.02
	Joint+MLM	1,952	0.82 ± 0.02	0.93 ± 0.01	0.68 ± 0.01	0.85 ± 0.02	0.76 ± 0.03

Experiments: Results & Analysis

- Models trained on one variant don't parse the other variant very well
  - Linguistic differences; corpus/genre differences
- If a model sees both variants in training, it doesn't seem to have a problem parsing both variants during inference
  - Even when both variants are seen with ½ volume (though the performance drops vs. full volume jointly-trained model)
- MLM with other Nahuatl variant data doesn't seem to have much impact
- No significant difference between monolingual, joint, and joint+auxiliary task

Experiments: Results & Analysis

What about the "multi-variant" performance of these different models?

Experiments: Results & Analysis

What about the "multi-variant" performance of these different models?

i.e. How do they perform on an evaluation set containing both variants?

Experiments: Results & Analysis

What about the "multi-variant" performance of these different models?

i.e. How do they perform on an evaluation set containing both variants?

Seeing both variants in training >>>>

Exp.	N	Lemma	UPOS	Morph.	UAS	LAS
Joint	1,952	0.86 ± 0.01	0.94 ± 0.01	0.75 ± 0.02	0.85 ± 0.01	0.77 ± 0.01
Joint Adj.	976	0.83 ± 0.01	0.92 ± 0.01	0.67 ± 0.01	0.81 ± 0.01	0.72 ± 0.01
azz alone	1,134	0.76 ± 0.01	0.79 ± 0.01	0.64 ± 0.01	0.74 ± 0.01	0.59 ± 0.01
nhi alone	818	0.74 ± 0.03	0.80 ± 0.01	0.53 ± 0.02	0.75 ± 0.02	0.60 ± 0.02

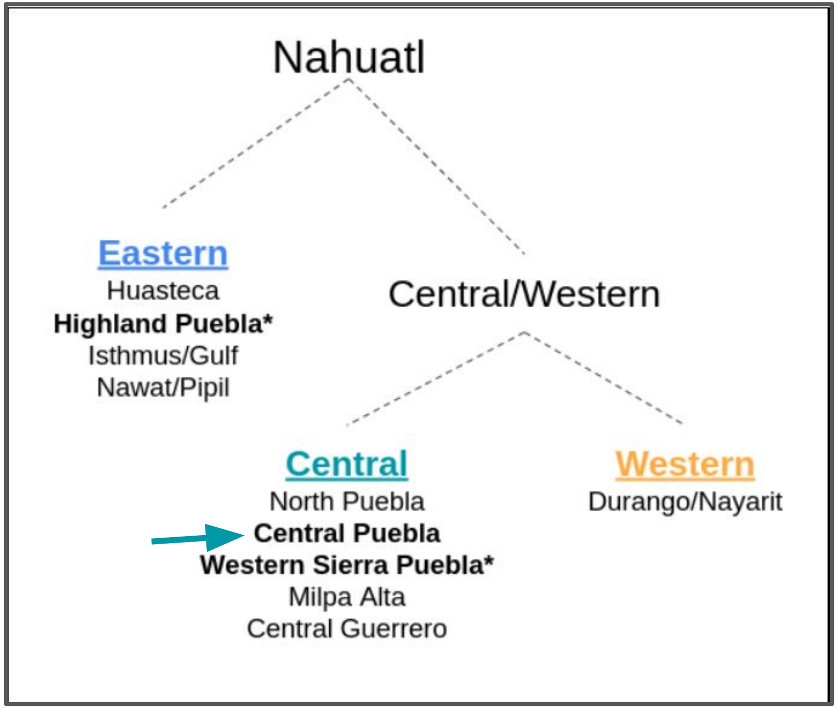
Experiments: Unseen variant (ncx)

Central Puebla Nahuatl (ncx)

- Central variant, also spoken in Puebla
- o- (**like nhi**), tl (**like nhi**), still has some short /e/ (**like azz**)

- *Casanova Stories* written by González-Casanova.

- Sampled 200 sentences, post-editing predicted parses to get the ground truth.



Experiments: Unseen variant (ncx)

Exp.	Lemma	UPOS	UAS	LAS
Joint	0.73	0.92	0.77	0.68
Joint Adj.	0.7	0.89	0.73	0.62
nhi alone	0.71	0.89	0.75	0.63
azz alone	0.62	0.64	0.63	0.36

Experiments: Unseen variant (ncx)

Data augmentation

- If azz underperforms due to linguistic distance, changing isogloss values in the azz data to match the "Central" values (i.e. "nhi-ifying" the azz data) should result in better performance.

- Likewise, toggling the isogloss values in nhi (i.e. "azz-ifying" the nhi data) should result in a model that is worse on the ncx evaluation data.

Prepend "o" to past-tense Verbs, and replace some "t"s with "tl"

`Amo niyekkochik tiotak` → `Amo oniyekkochik tiotlak`

Experiments: Unseen variant (ncx)

Exp.	Lemma	UPOS	UAS	LAS
Joint	0.73	0.92	0.77	0.68
Joint Adj.	0.7	0.89	0.73	0.62
nhi alone	0.71	0.89	0.75	0.63
azz-ified nhi	0.58	0.83	0.73	0.57
azz alone	0.62	0.64	0.63	0.36
nhi-ified azz	0.62	0.79	0.68	0.52

A Note about genre

- **Genre** seems to have an impact on **syntax** (Wang & Liu, 2017), e.g. adjacent dependency rates and dependency direction
- The two UD treebanks are not equivalent wrt. genre.
- This is a recognized confounding variable when analyzing these results as a function of dialectal differences only.

Future Work

- Continued efforts in corpus creation for Nahuatl variants

- Explore linguistically-informed data augmentation techniques in more depth

- Evaluate other auxiliary tasks and pre-training techniques for robust representations for Nahuatl
    - e.g. variant ID auxiliary task, pre-training with Spanish, MicroBERT (Gessler & Zeldes 2022).

- Empirically explore the impact on genre vs. linguistic features on cross-treebank parsing.

¡Tlasohcamatictzin!
¡Tasojkamatik!