Experiments in Multi-Variant Natural Language Processing for Nahuatl

Robert Pugh and **Francis M. Tyers**
Indiana University, Bloomington
Department of Linguistics
pughrob@iu.edu, ftyers@iu.edu

Abstract

Linguistic variation is a complicating factor for digital language technologies. This is particularly true for languages that lack an official "standard" variety, including many regional and minoritized languages. In this paper, we describe a set of experiments focused on multi-variant natural language processing for Nahuatl, an indigenous Mexican language with a high degree of linguistic variation and no single recognized standard variant. Using small (10k tokens), recently-published annotated datasets for two Nahuatl variants, we compare the performance of single-variant, cross-variant, and joint training, and explore how different models perform on a third Nahuatl variant, unseen in training. These results and the subsequent discussion contribute to efforts of developing low-resource NLP that is robust to diatopic variation. We share all code used to process the data and run the experiments.[1]

Figure 1: A map approximating the location of many of the Nahuatl variants spoken in Mexico. The colors correspond to the division defined in Kaufman (2001), blue for the Eastern branch, Turquoise for the Central branch, and Orange for the Western branch. We label the two variants for which we have training data in the form of UD treebanks. Importantly, this map is an approximation, and does not claim to represent every Nahuatl variant.

1 Introduction

Linguistic variation, though a ubiquitous feature of human language, is a complicating factor for digital language technologies. While natural language processing (NLP) has made significant advances in recent years, the "dialect gap," which refers to the drop in performance of NLP systems on non-standard linguistic varieties, remains (Kantharuban et al., 2023). In many cases, non-standard, low-resource variants are similar or related to a more uniform, standard variety with a greater number of linguistic resources. One popular approach to remedy this problem is to leverage a high-resource standard variant in concert with data augmentation methods to train models on a similar non-standard variant (Zampieri et al., 2020).

However, the case of a related, high-resource standard variant is not the only linguistic situation that speakers and writers of non-standard variants find themselves in. On the contrary, there are numerous distinct dialect situations across the world. In a treatment of such scenarios in Europe, Auer (2011) identifies a useful typology for thinking about the diversity of language situations with respect to standard languages and dialectal variation. Relevant to the present paper, this typology includes *exoglossic diglossia* or "Type 0", which describes a situation of multiple non-standard variants without any endoglossic standard. In these cases, if a standard variety does exist it is viewed as imported or significantly different from the vernacular dialects.

In the absence of a spoken or written standard variety ("Type 0"), in particular when there is little available annotated linguistic data for the non-standard varieties, developing digital language technologies robust to diatopic language variation is a particularly important and valuable objective.

[1] https://github.com/Lguyogiro/multidialectal-nlp-nahuatl

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 140–151

Nahuatl, a group of approximately 30 language varieties spoken in Mexico and Central America (Described in further detail in Section 2), fits the "Type 0" characterization quite well, given that there are a large number of recognized varieties and no single standard[2]. There also exists a vast body of literature in the language written primarily in historic Nahuatl varieties from the early colonial era, known as "Classical Nahuatl" (Gingerich; León-Portilla, 1985), to which speakers of contemporary Nahuatl varieties have little exposure.[3]

While these aspects of the Nahuatl language situation make it an interesting candidate for NLP research, they are not unique to Nahuatl. In fact, numerous indigenous language in Latin America fit the characterization of having many diatopically-diverse variants, no single contemporary standard, and a colonial-era written canon. Other examples include the Zapotec (Foreman and Lillehaugen, 2017; Flores-Marcial et al.; Hilts, 2003) and Quechuan (Luykx et al., 2016; Durston, 2008; Escobar, 2011) languages.

The present work evaluates a number of approaches to multi-variant NLP for Nahuatl. We leverage recently-published, relatively small Universal Dependencies corpora in two Nahuatl variants and compare monolingual model performance with that of cross-lingual and jointly-trained models, as well as the impact of leveraging multi-variant, unlabeled data by adding an auxiliary task during training.

Our goal in this effort is two-fold: (1) to set the stage for high-quality NLP models that support speakers of any variety of Nahuatl, leveraging their similarities, and (2) to inform similar efforts involving other languages in a similar dialect situation.

2 The Nahuatl Language Complex

Nahuatl is a polysynthetic, agglutinating Uto-Aztecan language spoken throughout Mexico and Mesoamerica. The Mexican Government's *Instituto Nacional de Lenguas Indígenas* (INALI) recognises 30 distinct Nahuatl varieties (INALI, 2009), with highly-variable levels of linguistic similarity and mutual intelligibility. Furthermore, linguistic similarity and mutual intelligibility is not always correlated with geographic distance, a fact that is due in part to multiple waves of migration of Nahuatl speakers leading speakers of different varieties to end up in close proximity to one another (Canger, 1988; Kaufman, 2001; Beekman and Christensen, 2003).

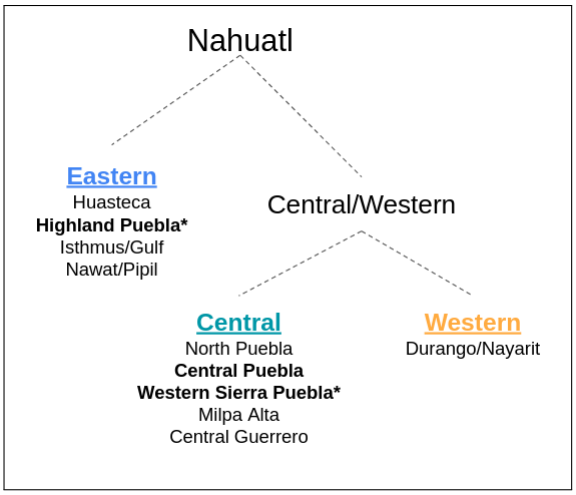

Figure 2: An abbreviated diagram of the sub-classification of Nahuatl variants, offering a glimpse at the taxonomic relationship between the variants we investigate here. The classification largely follows Kaufman (2001) using the same color-coding scheme in Figure 1. The variants used in this paper are bolded, and the two for which we have annotated training data are marked with an asterisk. The classification of the Central Guerrero variant follows (Lastra, 1986).

Dialectological research on Nahuatl dates back to at least (Lehmann, 1920). More recently, researchers largely converge around the dialect sub-classifications presented in (Lastra, 1986), (Canger, 1988), and (Kaufman, 2001) which, while not identical, agree on a number of important points, namely on the existence of Eastern Nahuatl varieties, which are thought to correspond to one wave of early migration, Central Nahuatl varieties, corresponding to the Nahuatl spoken in the valley of Mexico and in what is now Mexico City, and Western varieties, including Nayarit/Durango Nahuatl.

There is no unanimous consensus about the classification of Nahuatl variants, but for a number of cases there is widespread agreement (e.g. Pipil/Nawat of El Salvador and Sierra Puebla, or Highland Puebla, Nahuatl belonging to the Eastern group). Pharao Hansen (2014) provides some additional recommendations for the sub-classification between Eastern and Central/Western groups based on a survey of linguistic evidence. Nahuatl variants

[2]Alternatively, the label of "Pluricentric" (Clyne, 2012) may also be considered appropriate, though this typically refers to multiple standard, national languages, which is not the case of Nahuatl

[3]Interestingly, Sullivan (2011) describes a course with Nahuatl-speaking students focused on reading classical Nahuatl manuscripts, and notes that the students could read and understand it with little difficulty.

can differ at essentially every level of linguistic structure: Lexicon (e.g. *totoltetl* vs. *teksistli* "egg"), phonology (e.g. *e* vs. *i* (Canger and Dakin, 1985), *t–tl–l*, word-initial *e-* vs. *ye-*), morphology (e.g. the presence or absence of the "antecessive" *o-* for verbs in the past, the presence or absence of the perfective *-ki* suffix), and syntax (e.g. relative clauses (Pharao Hansen, 2015), and the order of certain adverbs with respect to verbs).

Additionally, since the invasion of Mexico in the 16th Century by the Spanish, Nahuatl has had close contact with Spanish, resulting in both in extensive "material borrowing" (Matras and Sakel, 2007) such as loanwords and new phonemes, but also a non-trivial amount of morphosyntactic "pattern borrowing" like syntactic calque, such as a development of the periphrastic future, and the development of adpositions from relational nouns (Farfán, 2008; Olko et al., 2018).

3 Related Work

Research on linguistic variation in NLP has recently become an important topic in the field, with now ten iterations of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) (Scherrer et al., 2023), which has over the years included a number of important and relevant shared tasks, such as similar language detection (Aepli et al., 2022) and cross-lingual parsing (Zampieri et al., 2017). Scherrer and Rambow (2010) explores approaches to NLP for the Swiss German dialect area that leverage geographic information, weighting rules using knowledge about the distribution of variant features in different regions. Also working on the Swiss German dialect continuum, Aepli (2018) evaluates syntactic parsing approaches including annotation projection, for which a parallel corpus with standard German was compiled, and delexicalized parsing.

These two approaches, annotation projection (Hwa et al., 2005; Agić et al., 2016) and delexicalized parsing (Zeman and Resnik, 2008), are common methods for cross-lingual parsing of related languages. More recently, the use of multilingual embedding representations used with neural network architectures has been shown to be quite effective for multilingual parsing (Ammar et al., 2016), particularly with pretrained transformer language models such as multilingual BERT (Devlin et al., 2019), as demonstrated in, e.g. in Kondratyuk and Straka (2019). Abdul-Mageed et al. (2021) build a language-specific transformer (with a large volume of data), reporting improved performance on multiple NLP tasks for a number of Arabic dialects.

One straightforward approach to multi-variant parsing is cross-lingual model transfer, wherein a model is trained on one variety (typically the higher-resource, standardized variety), and used on a different, related variety (Zampieri et al., 2020). Alternatively, work on two Norweigian standard languages, Bokmål and Nynorsk, found that simply combining the training data for closely-similar languages produces better results than straightforward model-transfer (Velldal et al., 2017).

While Nahuatl dialectology has a rich tradition in the field of linguistics (see Section 2), computational work addressing linguistic variation in Nahuatl is harder to come by. Efforts in this area include Farfan (2019)'s detailed analysis of the similarities of contemporary Nahuatl writing (from multiple variants) with Classical Nahuatl using a finite-state morphological analyzer built for the latter language, and Pugh and Tyers (2021), which found that simple, character-based language models, when evaluated across variants, track well with variant groupings and mutual intelligibility.

4 Data

We use recently published, linguistically-annotated datasets for two Nahuatl varieties: Highland Puebla Nahuatl (alternatively Sierra Puebla Nahuat, ISO-639: `azz`) (Pugh and Tyers, 2024) and Western Sierra Puebla Nahuatl (alternatively Zacatlán-Ahuacatlán-Tepetzintla Nahuatl, ISO-639: `nhi`) (Pugh et al., 2022), both spoken in the *Sierra Norte* region of the state of Puebla. Each of these datasets contains approximately 10,000 tokens, annotated using the Universal Dependencies (UD) (Nivre et al., 2020) framework for multiple levels of analysis: lemmatization, part-of-speech tagging, morphological analysis, and syntactic parsing.[4]

With respect to dialectal classification, Highland Puebla Nahuatl is clearly identifiable as an Eastern Nahuatl variety, and its place within the Nahuatl sub-classification is generally agreed-upon in the literature. Western Sierra Nahuatl's place is a bit trickier, in that it has a number of Central isoglosses, but also shares some features with the Eastern varieties (e.g. having /i/ where central vari-

[4] A quantitative comparison of the two treebanks can be found in Pugh and Tyers (2024).

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

Table 1: A breakdown of the datasets used in the paper and their total sizes. For the treebanks, which make up the data for the bulk of the experiments, we divided up the dataset 10 times into 90/10 splits in order to perform 10-fold cross validation. The variant labels listed with the Axolotl corpus are approximations based on an analysis of the 30 text sources that the sentences come from. The "Casanova stories" is a sample of texts from a larger collection, generously provided Joe Campbell.

eties have /e/). (Sasaki, 2015) provides a detailed comparison of Nahuatl variants spoken in Puebla's Northern Sierra, including Highland Puebla and Western Sierra Nahuatl. Table 2 provides an example of the differences between the variants.

It is worth noting that, though they are relatively distinct genealogically, these two variants are spoken in some adjacent communities and are in contact in areas such as Tetela de Ocampo, an azz-speaking municipality where some nhi-speakers go for commerce and school. It is therefore possible that these two varieties have more common features than any random selection of two variants. That being said, the two variants are distinguished by multiple isoglosses, e.g. the /t/-/tl/ distinction and use of the antecessive /o-/ in the past tense.

The nhi corpus consists of a combination of short stories, personal narratives, and grammar examples, and contains samples representing some linguistic diversity within the variant group (see Pugh et al. (2022) for specifics). The azz corpus, on the other hand, is more homogeneous, with the majority of the data coming from a single town and being largely of a single genre, namely descriptions of plants and their medicinal/culinary use.

For one experiment, we supplement the UD tree data with unlabeled Nahuatl text from the Axolotl corpus (Gutierrez-Vasques et al., 2016), a Nahuatl-Spanish parallel corpus with over 10k sentences.

Finally, we collect and annotate a small sample (about 2k tokens) of texts from the Central Puebla Nahuatl (ISO-639: ncx), a Central Nahuatl variety. The sample ("Casanova stories") is taken from a collection of short stories from Gonzalez-Casanova and prepared by Joe Campbell. We annotate the sample with the UD schema, but ignore morphological analyses due to the time-intensive nature of such annotation. This small dataset is used to evaluate our models' performances on a Nahuatl variety not seen during training.

4.1 Orthography

Numerous orthographic standards have been proposed over the years for written Nahuatl (using the Latin alphabet), but there is no real consensus. Often, written Nahuatl may be in a one-off orthography, and not necessarily consistent within a given text. Our data represents a variety of orthographies, and we normalize it using a finite-state transducer from the Py-Elotl Python package[5]. As the target orthography, we use one of the norms proposed by the Summer Institute of Linguistics (SIL) for Nahuatl, which uses 's' for /s/, 'c/qu' for /k/, 'tz' for /ts/, and 'u' for /w/. This decision is largely arbitrary. our motivation for choosing this instead of, for example, the INALI standard orthography (INALI, 2018), is the former's greater similarity to Spanish spellings (e.g. the graphemes "w" and "k" in Spanish are seen primarily only in loanwords). Since Nahuatl texts typically contain many Spanish words, and given the fact that the multilingual BERT model we use in our experiments was trained on a large amount of Spanish data, we chose to use an orthography that reflects Spanish spellings in order to better leverage the representations in the BERT model[6]. We use the normalized forms in all of the experiments in order to remove orthographic variation as a variable.

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

[6] Another option that would achieve the same goal would have been the ACK orthography, the only difference being the latter's lack of "s", which is relatively common in contemporary Spanish orthography. The quantification of orthographic similarity, and the extent to which orthography plays a role in Nahuatl parser performance using multilingual pretrained language models is a topic that we leave for future work.

azz	nhi	en
T**e**pos **teyin** tepaleuia mah **ica** se quita **teyin** amo u**e**li se quita ica se ixtololo.	T**i**pos**tl tlen** tepaleuia **ica** mo sequita **tlen** amo u**i**li 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."

Table 2: Example of two parallel sentences in azz and nhi. The azz text was taken from the corresponding treebank, and was translated by a speaker of nhi. Some specific differences are bolded, and include the raising of short /e/ in azz to /i/ in nhi, the *tl-t* isogloss, the absence of the antecessive *o-* on past tense verbs in azz, and a word-order difference with respect to the relational noun *ica* "with (instrumental)". The differences described here are by no means exhaustive.

Train	Eval	OOV%
nhi	nhi	$38\% \pm 3$
	azz	$81\% \pm 1$
	ncx	80%
azz	nhi	$83\% \pm 1$
	azz	$31\% \pm 3$
	ncx	87%
nhi + azz	nhi	$37\% \pm 3$
	azz	$30\% \pm 2$
	ncx	76%

Table 3: The percentage of out-of-vocabulary (OOV) tokens for the experiment configurations. When the Eval variant is nhi or azz, the experiments involve 10-fold cross-validation, so we average the OOV percentages over the folds and include the standard deviation. When calculating OOV percentage for the ncx data, we use the first fold of the training data. These numbers help give an initial impression of the difficulty of the different parsing tasks. Specifically, we see that, unsurprisingly, other-variant Eval datasets have substantially higher OOV percentages than same-variant Eval data.

5 Experiments

For all of the experiments described in this section, we use the MaChAmp toolkit (van der Goot et al., 2021) to fine-tune contextual subword embeddings from the pretrained multilingual BERT (mBERT) model[7] 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). During training, the best model is selected by summing the accuracy metrics of these tasks.

Due to the relatively low total volume of labeled data, we report results of 10-fold cross-validation.

5.1 Monolingual

We first evaluate the monolingual ("Mono" in Table 4), i.e. single variant, performance of the two Nahuatl variants, which serves as a benchmark for comparison with subsequent models. Intuitively, we expect these models to perform best on their respective variants, but be less robust when faced with multi-variant data.

5.2 Cross-Variant

Secondly, in order to get a sense of how challenging multi-variant NLP actually is for Nahuatl, we test zero-shot, cross-variant model transfer ("Cross" in Table 4, i.e. training on one variant and evaluating on the other. The motivation behind this experiment is the recognition that, it could be the case that many Nahuatl variants are similar enough to one another that there is no real need for special efforts targeted at multi-variant NLP for the language. If this were the case, we would expect zero-shot, cross-variant performance to be comparable with that of a monolingual model.

Recognizing that a major limitation of our dataset is the fact that it only represents two out of 30 Nahuatl variants, we annotated a small sample of short stories in a third variant, Central Puebla variety (ncx). We evaluate zero-shot, cross-lingual experiments on this dataset, as well as the performance of models jointly trained on both nhi and azz training sets. The objective of this experiment is to provide better a sense of the multi-variant capabilities of a model trained on limited data representing only a small set of Nahuatl varieties.

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

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

Table 4: Accuracy of a neural, multi-task UD parsing model in various training configurations. Each result is the average performance over 10 folds, followed by the standard deviation of the performance distribution. Note that, given the distribution overlap, not much can be said about the difference in performance of most of these experiments with the exception of the the cross-variant experiments, which consistently under-perform both monolingual (single-variant) and jointly trained models. Mono=Monolingual; Cross=Cross-variant (e.g. train on azz and predict on nhi); Joint=trained on the concatenation of both variants' corpora; Joint Adj.=like the Joint model, but only use half of the data from each variant during training; Joint w/ MLM=same as Joint, but with an additional masked language modeling task. "N" is the number of sentences in the training data for each experiment.

5.3 Joint Training

We train a model on the concatenation of the training data from the two Nahuatl variants, and evaluate its performance on each individual variant's evaluation data ("Joint" in Table 4). Ideally, given sufficient training data, the model can learn to implicitly detect the variant of an input text and, since a single set of model parameters is used for both variants, benefit from the similarities and increased coverage of Nahuatl linguistic features. Alternatively, it is plausible that the diatopic variation could add unhelpful noise during training.

By combining the training sets from two variants, we are also in effect doubling the training data size. To get a sense of how variant diversity in training effects model performance, while controlling for training data volume, we also experiment with combining just half of each of the nhi and azz training sets (*Joint Adj.* in Table 4).

5.4 Adding an Auxiliary Task During Training

We have emphasized that there is little available annotated Nahuatl text. However, there is a sizable amount of unlabeled text available that we can leverage to potentially improve system performance. We experiment with the Axolotl corpus (Gutierrez-Vasques et al., 2016), a parallel (Nahuatl-Spanish) collection of over 10,000 sentences of Nahuatl from multiple regions and time periods, including a large volume of colonial-era Classical Nahuatl.

Specifically, we perform the same multi-task approach described above, with an additional masked language modeling task using the Axolotl data ("Joint+MLM" in Table 4).

Since part of the azz treebank comes from the Axolotl corpus, we remove all text from source before creating this datasets in order to avoid data leakage.

6 Results and Discussion

The results of our experiments can be seen in Table 4. All results report the average and standard deviation of the performance on 10 folds.

6.1 Monolingual and Cross-variant Performance

Comparing the monolingual model performances, we note that the azz model performs either the same or better than the nhi model on nearly every task. This is likely due to the aforementioned greater homogeneity of the linguistic samples and genre in the azz treebank.

Secondly, the performance drops significantly from the monolingual models to the cross-variant models. Given the linguistic differences between the two variants, not to mention other differing characteristics of the corpora, this is largely expected. These results suggest the importance of focusing

on developing multi-variant capabilities in Nahuatl NLP, since these data appear to be different enough to impede straightforward cross-variant transfer, at least with small data volume. Upon collecting more annotated data, it would be valuable to also evaluate the monolingual models on same-variant, different-genre data in order to tease apart the influence of linguistic variation and other sources of variation in the corpora.

6.2 Analyzing Multi-Variant Performance

In analyzing the results of these experiments, we are most interested in the *multi-variant* performance. For "Joint" experiments, where the training data of both training variants is concatenated, the multi-variant performance is the combined performance on both variants. These models can be compared with a monolingual model evaluated on both variants (e.g. the monolingual result on `azz` and the cross-variant result on `nhi`).

For both variants, the jointly-trained model (See the "Joint" rows in Table 4) performs on par with two respective monolingual models, despite not having explicit language labels. For some tasks, the jointly-trained model has a higher average performance (taking error into consideration, however, the difference is not robust).

While the high performance of the jointly-trained model may be due to exposing the model to linguistic diversity during training, an important caveat is that the jointly-trained experiment has twice the volume of training data as the monolingual or cross-variant experiments. In order to investigate the extent to which data volume alone (versus, e.g. more robust learning during training) can explain the good multi-variant performance of the jointly-trained model, we also performed a volume-adjusted joint training experiment by combining half of the training set from each variant.

The results of this experiment and a comparison with the full jointly-trained model, are listed in Table 5. Unsurprisingly, here we see a dip in performance compared to the full jointly-trained model. Nonetheless, the volume-adjusted jointly-trained model still shows better multi-variant performance than the monolingual equivalent (monolingual cross-variant), supporting the utility of diverse training data.

6.3 The Effect of an Additional Training Task

Even for the model trained on the concatenation of datasets, the total available training data volume of is low (barely over 2k sentences) compared to so-called "high-resource" scenarios. Since no Nahuatl variant nor any genetically- or aerially-related language (with the exception of perhaps Spanish) was included in the multilingual BERT training data, we are interested in how we might be able to use additional unlabeled Nahuatl data, even if from different varieties or time periods, to improve the mBERT representations for Nahuatl.

We investigated whether training on an additional task, MLM using the Axolotl data, improves the model's Nahuatl representations, impacting parser performance. However, we do not see a significant impact: results for all tasks were still within the estimated margin of error (1 standard deviation of the 10-fold results) when compared to the jointly-trained model with no auxiliary tasks.

7 Performance of Monolingual and Multi-Variant Models on a Third, Unseen Nahuatl Variant

We evaluate the different trained models on parsing text from the unseen `ncx` variant. Performance on this unseen variant text are reported in Table 6.

As with the two-variant experiments listed in 4, the jointly-trained model, which is trained on the concatenation of the full `nhi` training data and the full `azz` training data, achieves the top performance on all tasks. Unlike the two-variant experiments, however, here the volume-adjusted jointly-trained model (trained on half of the `nhi` training data concatenated with half of the `azz` training data) does not out-perform both monolingual models. Instead, we see that the monolingual model trained on `nhi` data performs comparably to the volume-adjusted joint model on all tasks.

One plausible explanation is that the differences in performance between the two mono-variant models is due to a combination of variant similarity and genre overlap in the corpora. Namely, since *nhi* and *ncx* are both Central Nahuatl variants, they share a number of linguistic features, such as the presence of the /tl/ phoneme and the use of the antecessive *o-* on verbs in the past tense. For example, both *nhi* and *ncx* tokenize the antecessive suffix *o-* and tag it as AUX (e.g. *o niquitac*, "I saw it", which in `azz` is just *niquitac*). The *azz* corpus does not have any instances of this, since this variety does not mark past tense verbs with the antecessive, and instead the only instances of the word *o* are the Spanish conjunction meaning "or". As a result, the anteces-

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

Table 5: Comparing the multi-variant performance of different training configurations. The "azz and nhi alone" experiments use a monolingual model to parse multi-variant evaluation data. The "Joint" experiment trains a model on the concatenation of nhi and azz training data, leading to twice the training data volume as the other experiments. The "Joint Adj." experiment similarly trains on multi-variant data, but subsamples data from each variant to control for the possibility of data volume in and of itself being responsible for improved performance. "N" is the number of sentences in the training data for each experiment.

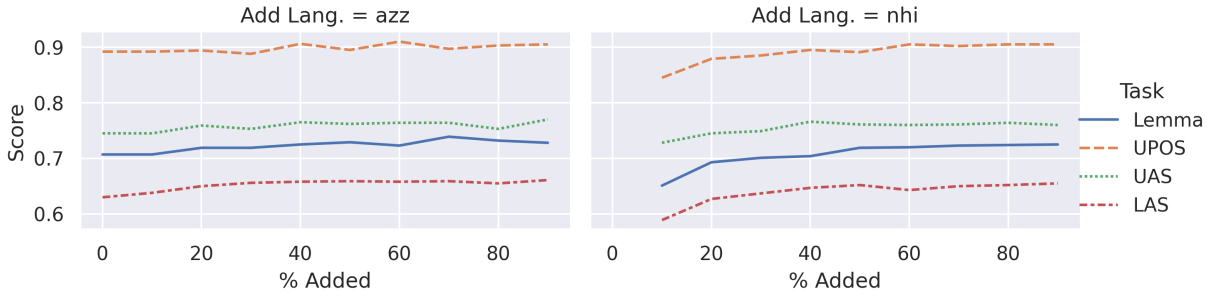

Figure 3: A plot of how the performance on the ncx data changes for the different tasks changes as we move from a monolingual model to a jointly-trained multi-variant model. As noted, ncx is linguistically-similar to nhi, and as such, adding azz data (the left plot) provides very minimal improvements, most of which seem to happen only once we've added 50% of the azz data. In the right plot, we see a larger improvement by just adding a small amount (the biggest marginal improvement happens when going from 0% to 20%) of nhi data to the azz-trained model.

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

Table 6: Performance of different models on an unseen Nahuatl variant, Central Puebla Nahuatl (ncx). Due to time constraints, we did not annotate the morphological analyses in this data, and thus do not report the performance.

sive in the ncx data is never correctly analyzed by the azz-only model.

To approach a better understanding of this proposed explanation of the results, we make copies of the monolingual datasets, altering the word forms with respect to both the antecessive *o-* and the /t/-/tl/ isogloss. That is, we make the nhi data more azz-like by removing the antecessive tokens and converting all instances of "tl" to "t". We expect a model trained on this version of the data to underperform on the ncx data since there is now a discrepancy in two prominent isogloss values. Likewise, we alter the azz add the antecessive token to all verbs with the morphological feature Tense=Past, and replace "t" with "tl" in positions that correspond to the latter segment in Nahuatl in general.[8] We expect a model trained on this dataset to perform better on the ncx data than the real azz data, since it has more common dialectal isoglosses.

The results show that the monolingual model trained on the nhi-ified azz data does indeed perform quite a bit better than that trained on the original azz data. Likewise, the model trained on the azz-ified nhi data peforms worse than that trained on the original nhi data. This shows the importance of dialectal similarity, even in the form of a pair of simple isoglosses. This result, while intuitive, is instructive for future work, since it indicates that, in the absence of more training data, variant-based

[8]The process of converting "t" to "tl" in the wordforms and lemmas was done via manual annotation.

data augmentation may be effective in increasing system performance.

It is also worth noting that, even after changing the isogloss values in the two datasets, the model trained on nhi data still outperforms that trained on the azz data when evaluating on the held-out Central variety, ncx. This fact indicates that morphological and syntactic factors are also at play. Furthermore, we also recognize the possible influence of genre on the performance differences.

With respect to genre and style, the unseen ncx text, a pair of short stories, more closely reflects the nhi corpus, which itself is largely made up of short stories, whereas the azz corpus consists almost entirely of transcriptions of recorded monologues describing the medicinal and culinary uses of plants. Findings such as those by Wang and Liu (2017), that a small but significant effect of genre on syntactic patterns such as adjacent dependency rate and dependency direction, may partially explain the much lower UAS and LAS performance by the azz model.

7.1 Learning Curve Experiment

To get a better sense of how adding different-variant data changes model performance on the ncx evaluation set, we perform a learning curve experiment for each variant, progressively adding 10% of the other variant's training data. The results of this experiment can be seen in Figure 3, plotting how the performance changes as we transition from a monolingual to a jointly-trained model by randomly adding data from the other variant. The azz model improves substantially with the addition of just a small amount (20%) of nhi data, and continues to improve as more data is added. The nhi-only model, on the other hand, improves only gradually with the addition of azz.

8 Future work

Revisiting the map in Figure 1, where we see that only two of Nahuatl variants have annotated treebanks, we recommend that the top priority for developing multi-variant NLP for Nahuatl be the continued collection of annotated corpora in additional variants and from diverse domains. Once more data is made available, we plan to empirically investigate the role of linguistic and genre similarity in multi-variant parsing using a variety of similarity metrics. For example, with an annotated test set for an additional Eastern Nahuatl variant, such as Huasteca Nahuatl, or azz sentences from a more diverse set of genres, further experiments could help shed light on the relative impact of genre and variant.

We also hope to explore other approaches to pretraining/auxiliary tasks in order to improve multi-variant parsing, such as building a language-specific pretrained model as described in Gessler and Zeldes (2022).

Finally, Nahuatl's long-standing contact with Spanish, a language with a significant number of annotated resources, offers a promising avenue of investigation of the extent to which Spanish data can be leveraged to improve NLP performance for Nahuatl.

9 Concluding Remarks

We reported the results of a series of experiments on UD parsing for Nahuatl, with a specific emphasis on multi-variant capabilities. We found, perhaps unsurprisingly, that the more examples of a given variant there are in the training data, the better the resulting model can perform on that variant. The multi-variant model performed as well as or better than two separate monolingual models, suggesting that having more data from diverse variants leads to a more robust model. Interestingly, we also found that a model's performance can be improved by superficially altering other-variant training data based on Nahuatl isoglosses. Though a number of points are still left to be investigated more thoroughly, this report serves as a first in-depth exploration of shallow Nahuatl NLP with the currently available datasets.

Acknowledgments

We are very grateful to Joe Campbell for his permission to use the Casanova stories, and to the anonymous reviewers for their helpful feedback.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Noëmi Aepli. 2018. *Parsing approaches for swiss german*. Ph.D. thesis, University of Zurich.

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Peter Auer. 2011. *26 Dialect vs. standard: a typology of scenarios in Europe*, pages 485–500. De Gruyter Mouton, Berlin, Boston.

Christopher S Beekman and Alexander F Christensen. 2003. Controlling for doubt and uncertainty through multiple lines of evidence: A new look at the mesoamerican nahua migrations. *Journal of Archaeological Method and Theory*, 10:111–164.

Una Canger. 1988. Subgrupos de los dialectos nahuas (1988). In J. Kathryn Josserand and Karen Dakin, editors, *Smoke and Mist: Mesoamerican Studies in Memory of Thelma D. Sullivan.Part. Oxford: BAR International Series 402 (Ii)*, volume 402 of *BAR lnternational*, pages 473–98. BAR, Oxford.

Una Canger and Karen Dakin. 1985. An inconspicuous basic split in nahuatl. *International Journal of American Linguistics*, 51(4):358–361.

Michael Clyne. 2012. *Pluricentric languages: Differing norms in different nations*, volume 62. Walter de Gruyter.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Durston. 2008. Native-language literacy in Colonial Peru: The question of mundane Quechua writing revisited. *Hispanic American Historical Review*, 88(1):41–70.

Anna María Escobar. 2011. Spanish in contact with Quechua. *The handbook of Hispanic sociolinguistics*, pages 321–352.

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

José Antonio Flores Farfán. 2008. *The Hispanisation of modern Nahuatl varieties*, pages 27–48. De Gruyter Mouton, Berlin, New York.

Xóchitl Flores-Marcial, Moisés García Guzmán, Felipe H. Lopez, George Aaron Broadwell, Alejandra Dubcovsky, May Helena Plumb, Mike Zarafonetis, and Brook Danielle Lillehaugen. *Caseidyneën Saën – Learning Together: Colonial Valley Zapotec Teaching Materials*.

John Foreman and Brook Danielle Lillehaugen. 2017. Positional verbs in colonial valley zapotec. *International Journal of American Linguistics*, 83(2):263–305.

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.

Luke Gessler and Amir Zeldes. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In *Proceedings of the The 2nd Workshop on Multilingual Representation Learning (MRL)*, pages 86–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Willard P Gingerich. A bibliographic introduction to twenty manuscripts of classical nahuatl literature. *Latin American Research Review*, 10(1):105–125.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 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.

Craig Hilts. 2003. From taxonomy to typology: The features of lexical contact phenomena in atepec zapotec-spanish linguistic contact.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara I. Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11:311 – 325.

INALI. 2009. *Catalogo De Las Lenguas Indigenas Nacionales: Variantes Linguisticas De Mexico Con Sus Autodenominaciones Y Referencias Geoestadisticas*. 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).

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. [to appear].

Terrence Kaufman. 2001. The history of the nawa language group from the earliest times to the sixteenth century: some initial results.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. *arXiv preprint arXiv:1904.02099*.

Yolanda Lastra. 1986. *Las areas dialectales del nahuatl moderno*. Universidad Nacional Autonoma de Mexico, Instituto de Investigaciones Antropologicas.

Walter Lehmann. 1920. Die sprachen zentral-amerikas in ihren beziehungen zueinander sowie zu sudamerika und mexiko, 1/2. *Zentral-Amerika, Teil I.*

Miguel León-Portilla. 1985. Nahuatl literature. In *Supplement to the Handbook of Middle American Indians, Volume 3: Literatures*, pages 7–43. University of Texas Press.

Aurolyn Luykx, Fernando García Rivera, and Félix Julca Guerrero. 2016. Communicative strategies across quechua languages. *International Journal of the Sociology of Language*, 2016(240):159–191.

Yaron Matras and Jeanette Sakel. 2007. Investigating the mechanisms of pattern replication in language convergence. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 31(4):829–865.

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.

Justyna Olko, Robert Borges, and John Sullivan. 2018. Convergence as the driving force of typological change in Nahuatl. *STUF-Language Typology and Universals*, 71(3):467–507.

Magnus Pharao Hansen. 2014. The East-West split in Nahuan dialectology: Reviewing the evidence and consolidating the grouping. In *Friends of Uto-Aztecan Workshop*.

Magnus Pharao Hansen. 2015. Dialectal variation in contemporary Nahuatl relative clause formation. AILS Seminar.

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, Marseille, France. European Language Resources Association.

Robert Pugh and Francis Tyers. 2021. Investigating variation in written forms of Nahuatl using character-based language models. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 21–27, Online. Association for Computational Linguistics.

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*.

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

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri. 2023. Tenth workshop on nlp for similar languages, varieties and dialects (vardial 2023). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*.

Yves Scherrer and Owen Rambow. 2010. Natural language processing for the swiss german dialect area. In *KONVENS*, pages 93–102.

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.

John Sullivan. 2011. The IDIEZ project: A model for indigenous language revitalization in higher education. *Collaborative Anthropologies*, 4(1):139–154.

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.

Erik Velldal, Lilja Øvrelid, and Petter Hohle. 2017. Joint UD Parsing of Norwegian Bokmål and Nynorsk. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 1–10.

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59:135–147.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*, pages 1–15.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.