Wav2pos: Exploring syntactic analysis from audio for Highland Puebla Nahuatl

Robert Pugh[♠] and **Varun Sreedhar**[♢] and **Francis Tyers**[♠]
pughrob@iu.edu, varunsreedhar14@gmail.com, ftyers@iu.edu
Indiana University, Bloomington
[♠]Department of Linguistics
[♢]Luddy School of Informatics, Computing, and Engineering

Abstract

We describe an approach to part-of-speech tagging from audio with very little human-annotated data, for Highland Puebla Nahuatl, a low-resource language of Mexico.[1] While automatic morphosyntactic analysis is typically trained on annotated textual data, large amounts of text is rarely available for low-resource, marginalized, and/or minority languages, and morphosyntactically-annotated data is even harder to come by. Much of the data from these languages may exist in the form of recordings, often only partially-transcribed or analyzed by field linguists working on language documentation projects. Given this relatively low-availability of text in the low-resource language scenario, we explore end-to-end automated morphosyntactic analysis directly from audio. The experiments described in this paper focus on one piece of morphosyntax, part-of-speech tagging, and builds on existing work in a high-resource setting. We use weak supervision to increase training volume, and explore a few techniques for generating word-level predictions from the acoustic features. Our experiments show promising results, despite less than 400 sentences of audio-aligned, manually-labeled text.

1 Introduction

Automatic morphosyntactic processing, such as morphological analysis or syntactic parsing, is an important task in Natural Language Processing (NLP) for the purposes of language documentation, feature-extraction for downstream NLP tasks (Sidorov, 2019; Wu et al., 2021; Sartakhti et al., 2021), and for quantitative corpus-based linguistic analysis (Tyers and Henderson, 2021; Kim et al., 2021).

The ample research exploring these tasks has, overwhelmingly, focused on textual data. However, text for low-resource, endangered, marginalized, and/or minority languages, which constitute a majority of the world's languages, is often sparsely available, if at all. Instead, much of the data from these languages may exist in the form of recordings, potentially only partially-transcribed or analyzed by field linguists working on language documentation projects. At the same time, recent progress in speech processing has resulted in powerful, pretrained speech representation models such as Wav2Vec2.0 (Baevski et al., 2020), which make it possible to achieve impressive ASR systems via fine-tuning on relatively little data (Yin et al., 2022). These same representations have also been shown to be useful in audio classification problems, such as speaker recognition (Vaessen and Van Leeuwen, 2022) and emotion detection (Pepino et al., 2021).

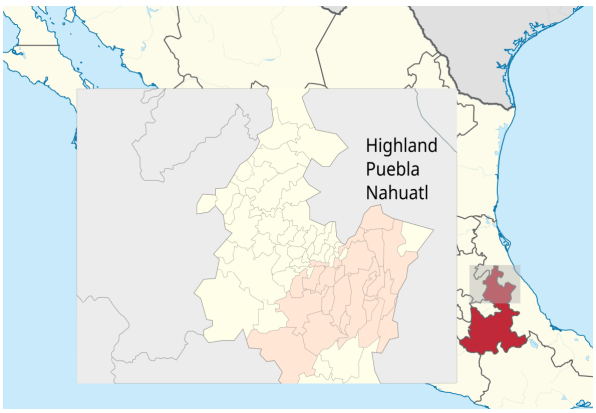

Figure 1: A map highlighting the 24 municipalities where HPN is spoken in the Sierra Norte de Puebla region of Mexico.

In the remainder of this paper, we explore end-to-end automated part-of-speech (POS) tagging directly from audio for an endangered Nahuatl variant, using a modest amount of transcribed data and only a few hundred sentences of annotated text. In light of a very limited data set.

[1]The code used for the work described here is available at https://github.com/VarunS9000/Wav2Pos

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 121–126

2 Highland Puebla Nahuatl

Nahuatl is a polysynthetic, agglutinating Uto-Aztecan language spoken throughout Mexico and Mesoamerica, made up of 30 recognized varieties (INALI, 2009).

Highland Puebla Nahuatl, (or *Sierra Puebla Nahuatl*, also referred to by INALI as *Náhuatl del noreste central*, `ISO-639-3` *azz*, henceforth HPN) is a Nahuatl variant group spoken by about 70,000 people (Ethnologue's 2007 estimate) in the Northeastern Sierra region of the state of Puebla, Mexico (see Figure 1) in 24 municipalities (INALI, 2009).

HPN has been the subject of documentary and descriptive linguistic efforts (Key, 1960; Robinson, 1970; Key and Key, 1953; Key and Richie de Key, 1953; Cortez Ocotlán, 2017). Furthermore it is one of two Nahuatl variants with a free and open morphosyntactically-annotated corpus, in the form of a Universal Dependencies treebank.

As an indigenous language of Mexico, HPN is considered at risk of being lost (INALI, 2012).

3 Related work

While most effort in the area of morphosyntactic analysis has focused on textual input, some recent work explores the idea of performing natural language processing directly from audio. Pupier et al. (2022) perform end-to-end dependency parsing for French from audio, by extracting audio features, aggregating them into audio word embeddings using LSTMs, and performing dependency parsing using these embeddings. These experiments used a dataset size consisting of 169,500 training sentences. Omachi et al. (2022) describe a non-autoregressive (non-transformer) method for performing end-to-end ASR and downstream NLP tasks such as named-entity recognition, performing part-of-speech tagging on a large.[2] corpus of spoken Japanese, and NER on a corpus of English containing approximately 10k training sentences (Bastianelli et al., 2020).

Shi et al. (2021) create a speech translation corpus of HPN using the same dataset as in the present paper, leveraging the fact that the entire dataset has transcriptions and translations (dataset details presented in Section 4).

[2]The size, e.g. number of sentences of words, of the corpus is not reported in Omachi et al. (2022) The publication presenting the corpus, Maekawa et al. (2000), describes it as having 7 million morphemes.

4 Data

The speech files, the transcriptions, and the pertinent metadata were obtained via the dataset from Amith et al. (2019) (hereafter "OpenSLR corpus"), which consists of about 50 hours of audio transcribed in ELAN. We processed the ELAN files, splitting the audio into utterance-level chunks using the Pydub Python library.[3]

The labeled HPN part-of-speech data comes from recently-released Highland Puebla Nahuatl UD treebank, which consists of (1) a small subset of the OpenSLR corpus annotated for morphosyntax, (2) a subset of texts in the *azz* variant from the multi-variant parallel corpus Axolotl (Gutierrez-Vasques et al., 2016), and (3) technical publications by the Sociedad Mexicana de Física.[4] Only (1), which contains 399 sentences and 3,463 tokens, has corresponding audio, and is held out for system evaluation. (2) and (3), totaling 838 sentences and 6,671 tokens, are used in training a simple text-based part-of-speech tagger, with which we generate synthetic data as described in Section 5.1.

The remaining OpenSLR corpus data (i.e. all of the audio/transcriptions that has not been annotated) is partitioned into a larger dataset for fine-tuning Wav2Vec2 (about 40k sentences), and a smaller dataset for training the audio-based POS tagger (about 7k sentences). We chose to use the majority of the data for Wav2Vec2 fine-tuning in order to ensure a high-performing ASR model since without that a POS tagger would have no words to align its POS tags with.[5] We chose not to use overlapping data for training both Wav2Vec2 and the Wav2pos tagger to avoid overfitting and poor performance on unseen data.

5 Methodology

In this section, we describe our method for training a POS tagger (Wav2pos) directly from audio with less than 400 labeled examples. This process involves training an acoustic feature extractor, generating silver training data, and aggregating the acoustic features to word-level.

[3]https://github.com/jiaaro/pydub

[4]https://site.inali.gob.mx/SMF/Libros2.0/nhtl/index.html

[5]In hindsight, it likely would have been worthwhile to experiment with different ways to partition this data, e.g. less data for Wav2Vec2 training and more for POS tagger training.

Dataset	Contents	Sentences	Tokens
Wav2Vec2 train*	OSLR − HPN	32k	285k
Text POS Tagger train	(HPN − OSLR) ∪ WSPN	1.7k	17.6k
Wav2pos train*	OSLR − HPN	8k	71k
Test data	HPN ∩ OSLR	363	2.4k

Table 1: A description of the contents of the different datasets. OSLR = OpenSLR data; HPN = Highland Puebla Nahuatl UD treebank; WSPN = Western Sierra Puebla Nahuatl UD Treebank. *The Wav2Vec2 fine-tuning data and the Wav2pos training data both come from the set of OpenSLR transcriptions not contained in the HPN treebank, but they are non-overlapping.

5.1 Synthetic label generation

The total amount of labeled POS data with corresponding audio for HPN is very small (363 sentences, 2k tokens). We hold it out for this purpose.

In order to produce enough labeled data to train the models, we label otherwise-unannotated OpenSLR transcriptions using a simple tagger. Specifically, we train an averaged perceptron model on the remaining UD trees (those sentences without corresponding audio), about 600 sentences. Since this is quite small for a training set, we supplement it with a UD treebank for another Nahuatl variant, Western Sierra Puebla Nahuatl (WSPN, `ISO-639-3` *nhi*) (Pugh et al., 2022), which added about 1k training sentences. The decision to add data from another variant is motivated by other recent work on Nahuatl syntactic parsing.

The averaged perceptron model uses words, substrings, previous words, and previous predicted tags as features. Once trained, we use it to predict POS tags on the unannotated OpenSLR transcriptions, resulting in "silver" training and validation data.

5.2 Extracting acoustic features

We use the unannotated (but transcribed) OpenSLR audio (split into a training and development set) to fine-tune the pret-trained Wav2Vec2.0 model (Baevski et al., 2020) on an ASR task. Our resulting fine-tuned model achieves a WER of 39% and CER of 18% on the held-out transcriptions.

In addition to being useful for automatically generating transcriptions, this fine-tuned model also gives us access to audio embeddings, corresponding to the discretized audio input, which have been fine-tuned for HPN. This sequence has many more elements than there are words (or even characters) in the sentence. We take two approaches to converting the longer sequence of acoustic embeddings into a single, word-level prediction in order to generate the part-of-speech tags.

5.3 Aggregating audio word embeddings with a BiLSTM

In the first Wav2pos approach, we first identify the subsequence of the acoustic embeddings by separating them by predicted whitespace characters. We pass each sequence (corresponding to segments of a single word) through a Long Short-Term Memory network (LSTM). The final hidden state is, then, a vector corresponding to a word. The sequence of word vectors is POS-tagged with a separate, bidirectional LSTM.[6]

5.4 Character-based prediction approach

As an alternative approach, we first reformat our data so that the label sequence, instead of consisting of a single POS tag per word, has a POS tag corresponding to each character (where the character's POS tag is that of its word). For example, for a transcription like *kemah niyas*, which originally is tagged `[INTJ, VERB]`, the label sequence is converted to `[INTJ, INTJ, INTJ, INTJ, INTJ, SPACE, VERB, VERB, VERB, VERB, VERB]`, such that each character in the transcription has a corresponding POS tag (note the inclusion of the `SPACE` tag corresponding to the word boundary). We pass the entire sequence of acoustic embeddings (without splitting them up into predicted audio words) through a BiLSTM, and make a POS prediction at each time step. For this approach, we use CTC loss in order to optimally-align the predicted POS tags with the labels. During inference, for a given word we choose the most frequent of its character-based POS tags as its tag.

5.5 Experiments

Given our silver training and validation data, and gold, human-annotated evaluation dataset, we compare the performance of three systems (as described in the previous section):

[6]The two LSTMs are trained jointly.

		Micro			Macro		
System	**Accuracy**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
apt	69.7	71.7	69.8	70.7	64.5	64.8	61.5
wb	53.2	57.1	53.2	55.1	51.4	48.2	46.3
cb	70.1	71.8	70.1	70.9	74.6	64.0	63.2

Table 2: Results comparing three approaches to POS-tagging our corpus. wb and cb correspond to systems that make predictions directly from audio, whereas the aptrepresents a pipeline system, wherein a text-based POS-tagger is run on the transcriptions output from the ASR system. While the two Wav2pos systems vary widely in their performance, the performance of the *cb* system suggests that the acoustic representations in the Wav2Vec2 model do in fact contain sufficient syntactic information.

apt: Averaged perceptron run on the output of the ASR. This method allows us to ascertain whether there is any benefit to calculating the POS tags directly from the audio instead of chaining the tagging with the ASR system.

wb: Word-based aggregation, where each hidden vector corresponding to an acoustic word (defined by the model's whitespace predictions) are first aggregated into a single word vector via an LSTM, and the sequence of aggregated word vectors is passed through a BiLSTM to predict the sequence of POS tags. This system is described above in Section 5.3

cb: Character-based aggregation, where the POS tag of a word is predicted for each of its characters, as described in Section 5.4.

5.6 Evaluation methodology

Since Wav2pos is based on the acoustic representations of the Wav2Vec2 model fine-tuned for ASR, and the ASR model may mistranscribe some words, our evaluation only takes into consideration words that the ASR model correctly transcribed. Specifically, we create tuples from the Wav2pos prediction and the ASR output, and match the ASR output to words in the correct transcription. If the word is transcribed correctly (i.e. it is found in the gold transcription), we compare the POS tags.

6 Results

The results of our three experiments are reported in Table 2. We note the passable performance of character-based Wav2pos model (cb), which is slightly better than the pipeline approach of tagging the transcriptions with a text-based POS-tagger (apt). This result suggests that indeed there is recoverable syntactic information represented in the acoustic feature embeddings learned by the Wav2Vec2 model, even (or especially) when these embeddings correspond to only a small piece of the word, as in the character model.

There is a significant difference in performance between the two Wav2pos models, the the wb model much worse on all metrics. While without more detailed analysis we can only speculate, it appears as though the aggregation step, which involves passing a sequence of acoustic vectors through an LSTM to produce a single "audio word vector," may introduce too many additional parameters for the model to learn given the relatively small amount of training data.

While these results are certainly interesting, they raise more questions than they answer, primarily as the result of the constrained set of experiments we performed. For future work, we plan to replicate these experiments, but using the same, larger Wav2Vec2 training dataset to train the Wav2pos models. We would also like to explore the hyperparameter space of these models in more depth, and try using a stronger text-based tagger, such as a multilingual pretrained transformer-based model, to create the silver data.

Finally, given the promising results for POS tagging, we are interested in expanding these efforts to other aspects of syntactic analysis such as dependency parsing.

7 Concluding remarks

We have presented a preliminary investigation of automated morphosyntactic analysis from audio with no human-labeled training data. We leveraged a large set of transcribed audio to fine-tune a Wav2Vec2 acoustic feature-extraction model, and experimented with producing POS-tags directly from the acoustic embeddings. We created our

training data by tagging unlabeled transcription data using a simple classifier model. The results showed that one of our audio-based POS-tagging models performed slightly better than using the text-based tagger to tag the transcriptions.

8 Acknowledgements

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. We also thank the anonymous reviewers for their helpful feedback.

References

Jonathan D. Amith, Amelia Dominguez Alcántara, Hermelindo Salazar Osollo, Ceferino Salgado Castañeda, and Eleuterio Gorostiza Salazar. 2019. Audio corpus of Sierra Nororiental and Sierra Norte de Puebla Nahuat(l) with accompanying time-code transcriptions in ELAN.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Pedro Cortez Ocotlán. 2017. *Diccionario Nahuat–Español de la Sierra Nororiental del Estado de Puebla*. Tetsijtsilin, Tzinacapan, Cuetzalan.

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.

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. 2012. *México: Lenguas indígenas nacionales en riesgo de desaparición*. Instituto Nacional de Lenguas Indigenas, México.

Harold Key. 1960. Stem construction and affixation of Sierra Nahuat verbs. *International Journal of American Linguistics*, 28(2):130–145.

Harold Key and Mary Richie de Key. 1953. *Vocabulario Mejicano de la Sierra de Zacapoaxtla, Puebla*. Instituto Lingüístico de Verano, México, D.F.

Mary Key and Harold Key. 1953. The phonemes of sierra nahuat. *International Journal of American Linguistics*, 19(1):53–56.

Jongin Kim, Nayoung Choi, Seunghyun Lim, Jungwhan Kim, Soojin Chung, Hyunsoo Woo, Min Song, and Jinho D. Choi. 2021. Analysis of zero-shot crosslingual learning between English and Korean for named entity recognition. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 224–237, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Motoi Omachi, Yuya Fujita, Shinji Watanabe, and Tianzi Wang. 2022. Non-autoregressive end-to-end automatic speech recognition incorporating downstream natural language processing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6772–6776.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

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.

Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, and Jérôme Goulian. 2022. End-to-end dependency parsing of spoken french. In *Interspeech*.

Dow F. Robinson. 1970. *Aztec studies 2: Sierra Nahuat word structure*. Summer Institute of Linguistics.

Moein Salimi Sartakhti, Romina Etezadi, and Mehrnoush Shamsfard. 2021. Improving Persian relation extraction models by data augmentation. In *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*, pages 32–37, Trento, Italy. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63, Online. Association for Computational Linguistics.

Grigori Sidorov. 2019. *Syntactic n-grams in computational linguistics*. Springer.

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.

Nik Vaessen and David A Van Leeuwen. 2022. Fine-tuning wav2vec2 for speaker recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7967–7971. IEEE.

Sixing Wu, Ying Li, Minghui Wang, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2021. More is better: Enhancing open-domain dialogue generation via multi-source heterogeneous knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2286–2300, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haiyan Yin, Dingcheng Li, and Ping Li. 2022. Learning to selectively learn for weakly supervised paraphrase generation with model-based reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1385–1395, Seattle, United States. Association for Computational Linguistics.