

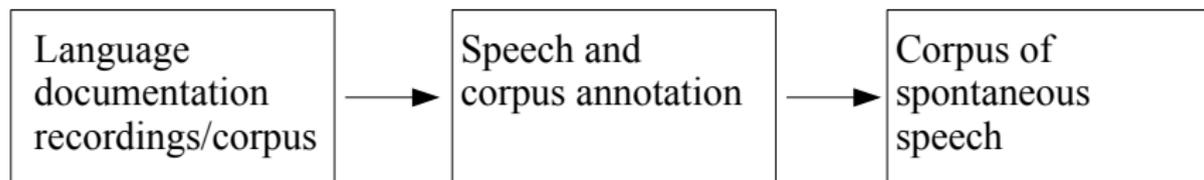
# Phonetics and DEL/DLI: experimental methods and tools for endangered language corpora

Christian DiCanio  
cdicanio@buffalo.edu

Department of Linguistics  
University at Buffalo

1/3/20

# The documentation research pipeline

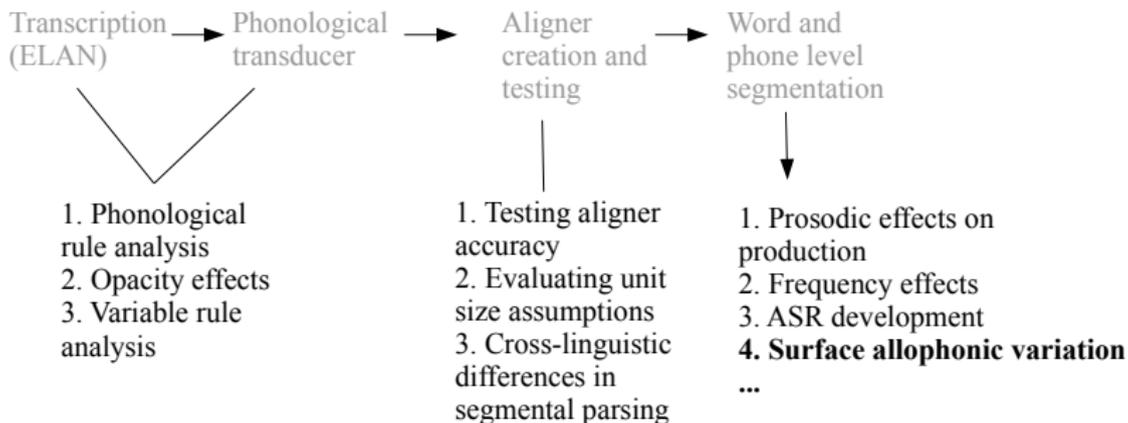


Opportunities for collaborative research exist **both** during the process of corpus annotation and after annotation completes.

# A rich pipeline for collaboration

Automatic tools aid in corpus annotation. Evaluating their utility is both technically useful and theoretically-relevant to linguistics as a discipline.

(Babinski et al., 2019; DiCanio et al., 2013; Johnson et al., 2018; Kirschenbaum et al., 2012; Michaud et al., 2018; Schreer and Schneider, 2012; Tang and Bennett, 2019)



# Corpus phonetics

Linguistic research is increasingly interested in examining speech in ecologically typical contexts in well-studied languages.

(Chodroff and Wilson, 2017; Davidson, 2016, 2018; Stuart-Smith et al., 2015; Wedel et al., 2013)

There is increasingly more research on the phonetics of endangered languages using spontaneous speech corpora.

(Coto-Solano, 2017; DiCanio et al., 2015; DiCanio and Whalen, 2015; Evans et al., 2008; Fletcher and Evans, 2002; Kakadelis, 2018; Kaland and Himmelmann, 2019; Muehlbauer, 2012; Tang and Bennett, 2018)

Phonetic variation is highly structured (Ladd, 2014) and investigating the phonetics of variation in endangered language corpora can shed light on scientific questions related to constraints on this structure.

# Phonetic case studies

- Corpus of Yoloxóchtitl Mixtec (Otomanguean: Mexico):

*Scientific question:* What explains variable lenition?

*Annotation/scientific question:* Can surface *phonetic* variation be predicted/modeled with deep neural networks?

- Corpus of Itunyoso Triqui (Otomanguean: Mexico):

*Annotation/scientific question:* How well does an aligner in a laryngeally-complex language work?

*Scientific question:* How do tones vary in spontaneous speech?

# I. The Yoloxóchitl Mixtec corpus

- Otomanguenan, spoken in Guerrero, Mexico (~2500 speakers).
- 120 hours of transcribed personal narratives, stories, and folklore; 30 speakers (Amith & Castillo García, 2009 – present).
- Phonological/phonetic fieldwork (Castillo García, 2007; DiCano et al., 2014, 2018, 2019; Palancar et al., 2016).



- Relatively simple consonant and vowel inventories; contrastive vowel nasalization (Castillo García, 2007; DiCanio et al., 2019).
- Content words are minimally bimoraic and no codas are permitted (Castillo García, 2007).
- Contrastive glottalization and a very complex tonal inventory; root-final stress realized via lengthening (DiCanio et al., 2018, 2019).

Melody	Word	Gloss	Melody	Word	Gloss
1.1	ta <sup>1</sup> ma <sup>1</sup>	<i>without appetite</i>	4.13	na <sup>4</sup> ma <sup>13</sup>	<i>is changing</i>
1.3	na <sup>1</sup> ma <sup>3</sup>	<i>to change (intr)</i>	4.14	nda <sup>4</sup> ta <sup>14</sup>	<i>is splitting up</i>
1.4	na <sup>1</sup> ma <sup>4</sup>	<i>soap</i>	4.24	ya <sup>4</sup> ma <sup>24</sup>	<i>Amuzgo person</i>
1.32	na <sup>1</sup> ma <sup>32</sup>	<i>I will change myself</i>	4.42	na <sup>4</sup> ma <sup>42</sup>	<i>I often pile rocks</i>
1.42	na <sup>1</sup> ma <sup>42</sup>	<i>my soap</i>	13.2	hi <sup>13</sup> ni <sup>2</sup>	<i>has seen</i>
3.2	na <sup>3</sup> ma <sup>2</sup>	<i>wall</i>	13.3	na <sup>13</sup> na <sup>3</sup>	<i>has photographed oneself</i>
3.3	na <sup>3</sup> ma <sup>3</sup>	<i>to change (tr)</i>	13.4	na <sup>13</sup> ma <sup>4</sup>	<i>has piled rocks</i>
3.4	na <sup>3</sup> ma <sup>4</sup>	<i>sprout</i>	14.2	na <sup>14</sup> ma <sup>2</sup>	<i>I will not change</i>
3.42	na <sup>3</sup> ma <sup>42</sup>	<i>I will pile rocks</i>	14.3	na <sup>14</sup> ma <sup>3</sup>	<i>to not change</i>
4.1	ka <sup>4</sup> nda <sup>1</sup>	<i>is moving (intr)</i>	14.4	na <sup>14</sup> ma <sup>4</sup>	<i>to not pile rocks</i>
4.2	na <sup>4</sup> ma <sup>2</sup>	<i>I am changing</i>	14.13	na <sup>14</sup> ma <sup>13</sup>	<i>to not change oneself</i>
4.3	na <sup>4</sup> ma <sup>3</sup>	<i>it is changing</i>	14.14	nda <sup>14</sup> ta <sup>14</sup>	<i>to not split up</i>
4.4	na <sup>4</sup> ma <sup>4</sup>	<i>is piling rocks</i>	14.42	na <sup>14</sup> ma <sup>42</sup>	<i>I will not pile rocks</i>

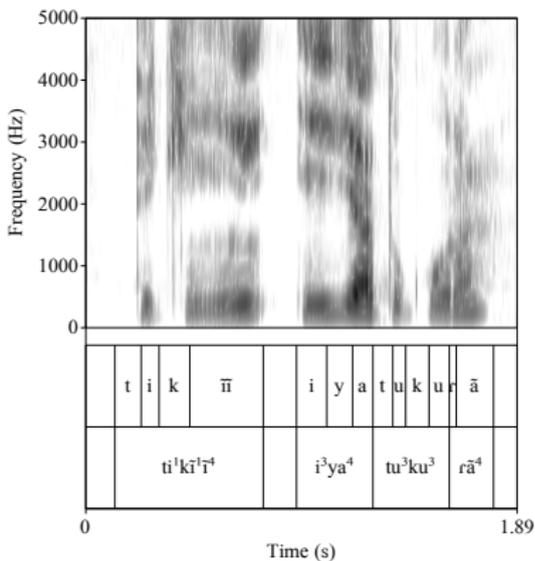
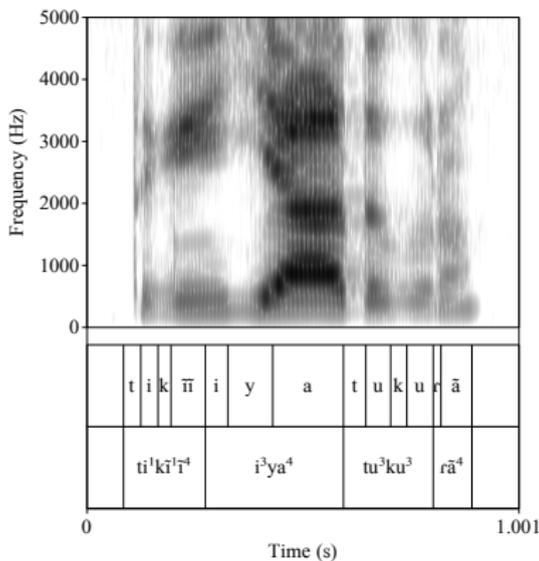
# The annotation pipeline for the YM corpus

- 1 Collection and transcription of 100+ hours of spontaneous speech via NSF DEL grant #0966462 (Amith, PI).
- 2 Creation of phonological transducer and forced alignment system via NSF DEL grant #1603323 (DiCano, PI).
- 3 ASR development via NSF DEL grants #1500738 and 1500595 (Kathol & Amith PIs). Outcome: Testing how much inclusion of tonal information in a complex tone language enhances performance (Mitra et al., 2016)

Outstanding issue: The transcription of the force aligned speech signal is only as accurate as the phonological rules specified in the transducer. Variable rules can not be applied.

# Consonant variability

[ti<sup>1</sup>yi<sup>1</sup>ri<sup>4</sup> ja<sup>4</sup> du<sup>3</sup>yu<sup>3</sup> rã<sup>4</sup>] (left) vs. [ti<sup>1</sup>ki<sup>1</sup>ri<sup>4</sup> i<sup>3</sup>ja<sup>4</sup> tu<sup>3</sup>ku<sup>3</sup> rã<sup>4</sup>] (right)



'...the sour tamale again, then.'

## Variable obstruent lenition

This lenition is not predictable by rule; stops always have closure in elicited speech (DiCanio et al., 2019).

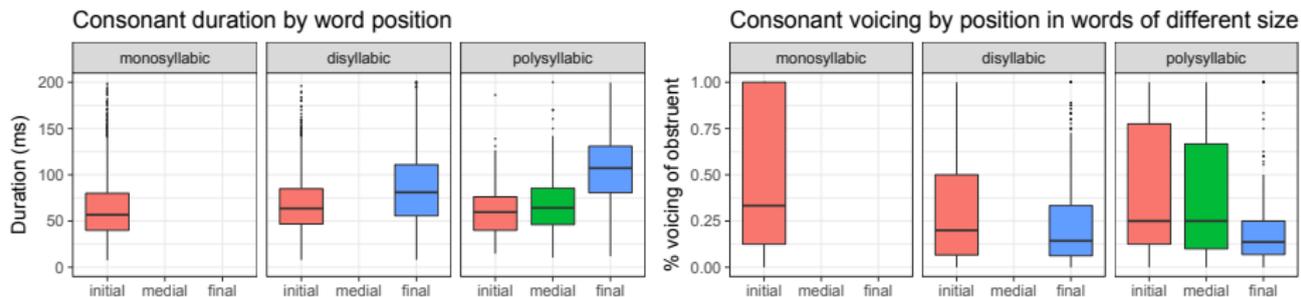
What predicts this lenition in Yoloxóchitl Mixtec? Can we model it?

Stops may be produced with variable degrees of lenition in spontaneous speech. (Bouavichith and Davidson, 2013; Davidson, 2011; Hualde et al., 2011; Katz, 2016; Katz and Fricke, 2018; Katz and Pitzanti, 2019; Lewis, 2001; Torreira and Ernestus, 2011; Warner and Tucker, 2011)

Across languages, manner/voicing lenition is more common in word-medial position and in the onsets of unstressed syllables than in word-initial position or in the onsets of stressed syllables. What about here?

# Scientific outcome

An examination of 107 minutes of force-aligned speech reveals that voicing lenition is more common in *word-initial* position than in word-medial position (DiCano et al., 2017).



Novel finding! Previous work argues that lenition is resisted at word boundaries (Katz and Fricke, 2018; Katz and Pitzanti, 2019).

## Can we model this lenition?

We examined 89 minutes of corpus used for voicing/lenition study and coded 4472 stop tokens (/t, k/) for lenition type.

	Vcls stop	Partially vcd stop	Voiced stop	Voiced fric.	Voiced approx.	Nasal	Tap	Deleted
/t/	17.9%	33.0%	21.2%	15.8%	2.7%	6.6%	1.2%	1.6%
/k/	15.3%	20.0%	16.4%	33.5%	7.9%	1.5%	NA	4.8%

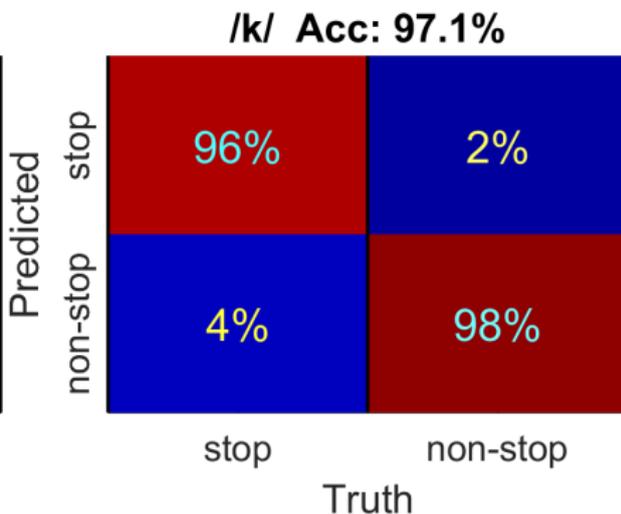
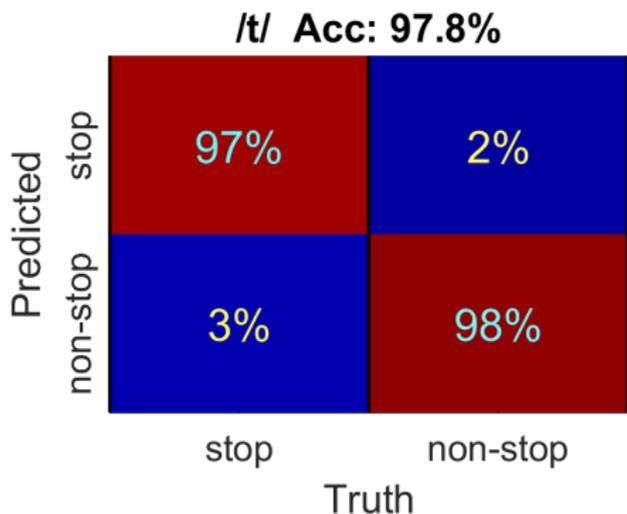
Predicting surface phonetic variation not only permits greater detail in the speech corpus, but allows one to examine low-level variation in speech production without needing to code the acoustic data by hand.

# Methods: DNN modelling

- We can use the allophonic labelling from the 4,472 stop tokens to train DNNs (Deep neural networks) to categorize surface phonetic allophones.
- Six models trained: 2-way, 3-way, 4-way models on /t/ and on /k/; (500 nrns) (Hinton et al., 2012).
- 20 MFCC coefficients extracted from each hanning-windowed (10 ms, 2ms step) acoustic signal (48 kHz > 16 kHz) for each stop token. MFCCs were standardized, normalized, and rescaled.
- Models trained on 80% of data, fine-tuned on 10% cross-validation set, and tested on remaining 10% (random split).

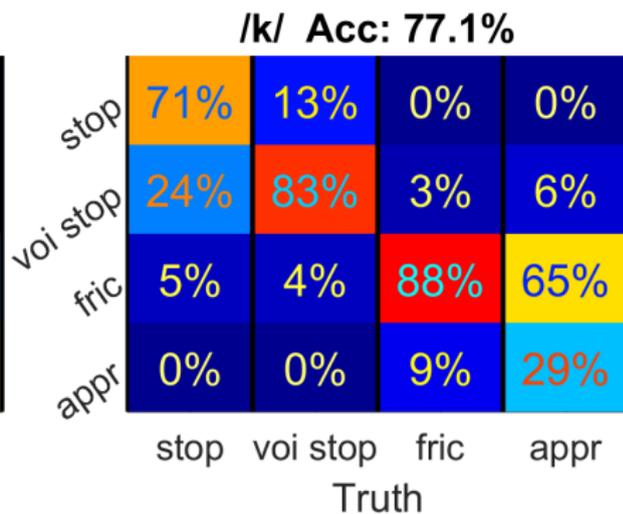
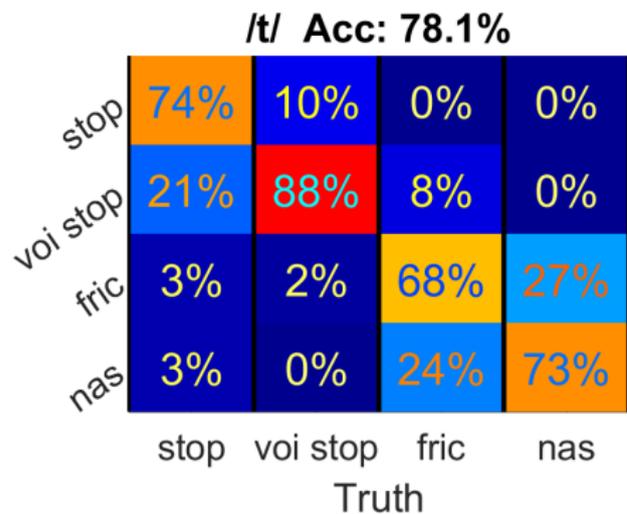
## 2-way categorization

High accuracy found – stop vs. non-stop



## 4-way categorization

Good accuracy found – voiceless stop vs. voiced stop vs. fricative vs. sonorant (nasal or approximant).



## Annotation/Scientific Outcome:

- DNN models can detect allophones from spontaneous speech despite limited training data.
- Excellent stop/continuant identification, though approximants were more poorly identified. The four-way model showed good performance in voiceless-voiced stop identification.
- Larger intellectual impacts:
  - Detection of stop/continuant distinction important for the diagnosis of childhood apraxia of speech (Davis et al., 1998).
  - Automatic detection of surface phonetic variation is generally relevant for questions in prosody, speech articulation, and sociolinguistics.

## II. Forced alignment

What about the step of segmenting the speech corpus? This is useful for dictionary work, usage-based linguistic analysis, discourse analysis, and other research areas.

Most alignment systems are trained on major languages, like English, Spanish, French, and Mandarin Chinese.

(Adda-Decker and Snoeren, 2011; Lin et al., 2005; Malfrère et al., 2003; Yuan and Liberman, 2008, 2009)

Creation of alignment systems has become easier with under-resourced languages.

e.g. in Gaelic (Ní Chasaide et al., 2006), Kaqchikel (Tang and Bennett, 2018), Tongan (Johnson et al., 2018), Uspanteko (Tang and Bennett, 2019), and Xhosa (Roux and Visagie, 2007).

# The Itunyoso Triqui Corpus

- Otomanguean, spoken in Oaxaca, Mexico (~2500 speakers).
- Running speech: 25 hours of transcribed personal narratives, stories, and folklore; 31 speakers, collected between 2013 - 2017.
- Initial transcription done by trained native speakers, subsequent revision with PI (DiCano).
- Phonological/phonetic fieldwork (DiCano, 2008, 2010, 2012c,b, 2016).

Challenges for alignment - laryngeal complexity and occasional code-switching with Spanish.

# Itunyoso Triqui phonology

(DiCanio, 2008, 2010, 2016)

Elaborate consonant inventory (glottalized sonorants, singleton-geminate contrast, coda glottal consonants).

Nine contrastive tones on root-final syllables.

Tone	Open syllable		Coda /h/		Coda /ʔ/	
	Word	Gloss	Word	Gloss	Word	Gloss
/4/	yū <sup>4</sup>	'earthquake'	yāh <sup>4</sup>	'dirt'	niʔ <sup>4</sup>	'see.1DU'
/3/	yū <sup>3</sup>	'palm leaf'	yāh <sup>3</sup>	'paper'	tsiʔ <sup>3</sup>	'pulque'
/2/	ū <sup>2</sup>	'nine'	tah <sup>2</sup>	'delicious'	ttʃiʔ <sup>2</sup>	'ten'
/1/	yū <sup>1</sup>	'loose'	kāh <sup>1</sup>	'naked'	tsiʔ <sup>1</sup>	'sweet'
/45/			toh <sup>45</sup>	'forehead'		
/13/	yo <sup>13</sup>	'fast (adj.)'	toh <sup>13</sup>	'a little'		
/43/	ra <sup>43</sup>	'want'	nnāh <sup>43</sup>	'mother!'		
/32/	rā <sup>32</sup>	'durable'	nnāh <sup>32</sup>	'cigarette'		
/31/	rā <sup>31</sup>	'lightning'				

# Collaboration - Creating an aligner for Triqui

Issue #1: Creation of pronunciation dictionary for Triqui interspersed with Spanish loanwords.

Issue #2: Vowel-glottal segmentation is generally bad, so we treated vowels with glottal codas as gestalts, e.g. /aʔ/ is a single phone, not /a/+/?/.

The screenshot shows an audio alignment interface with a waveform at the top and several layers of text and phonetic transcription below. The time axis ranges from 00:04:56.600 to 00:04:58.000.

Layers shown:

- Traducción [58]**: (empty)
- Translatio [58]**: (empty)
- CLG [138]**: yyaj13 roh4 ni2 ngo2yan2 ba32 ngwi31 cha1na1 ni2 ngo2yan2 ba32 si4sto43
- Traducción [138]**: Cuando ahora somos iguales las mujeres y son iguales los hombres
- Translatio [138]**: Because now we women are equal and we're equal to men.
- CLG\_phon [2660]**: yyaj13 roh4 ni2 ngo2yan2 ba32 ngwi31 cha1n
- CLG\_pho [6895]**: sp |j: |ah |r |oʔ |n |i~ |ng |o |j |a~ |B |a |ng |\_w|i |tS

# How to assess alignment

Table 1: *Accuracies at different tolerances (percentage below a cutoff) for absolute differences between force-aligned boundaries using MFA-LS aligner, and gold-standard annotations.*

	Tolerance (ms)			
	<10	<25	<50	<100
Word boundaries (Buckeye)	0.33	0.68	0.88	0.97
Phone boundaries (Buckeye)	0.41	0.77	0.93	0.98
Phone boundaries (Phonsay)	0.36	0.72	0.88	0.95

(McAuliffe et al., 2017)

## Annotation outcome: Assessing the Triqui aligner

Built an aligner on larger portion of the corpus and then tested it on 33.8 minutes of speech.

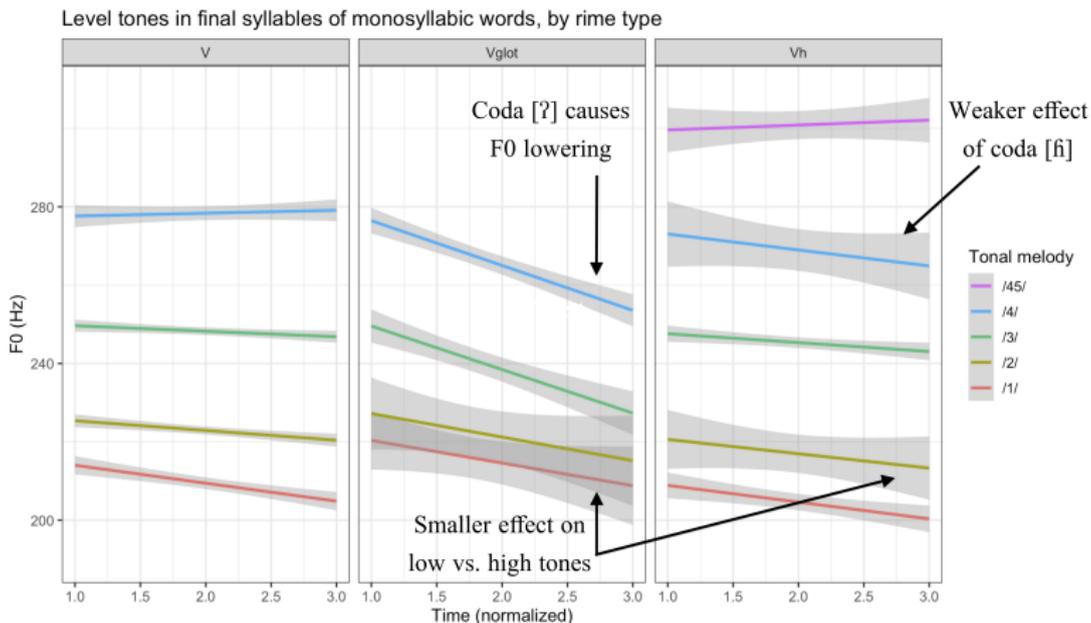
Tolerance	10 ms	20 ms	30 ms	40 ms	50 ms
% phones in corpus	46.7%	77.1%	89.2%	93.7%	95.9%

These results are noteworthy since the aligner is (a) partially multilingual and (b) manages to align glottalized segments (c.f. DiCanio et al. (2013)).

What can the alignment be used for? Tonal recognition benefits from production variability research (Lin et al., 2018).

# Scientific outcome - corpus tone production

Replication of (DiCano, 2012a) examining the effect of glottal consonants on tone with force-aligned spontaneous speech from 2 female speakers.



## Discussion/Conclusions

Scientific outcomes emerge both from the annotation process (DNN model, forced alignment model) and as a result of it (segmental lenition, tone production).

The phonological/phonetic annotation pipeline provides a unique set of collaborative opportunities for language documentarians, phoneticians/phonologists, and computationally-oriented researchers.

The outcomes of this collaboration can extend *beyond* each research area, more generally demonstrating the value of language documentation to scientific research.

# Acknowledgements - diverse collaborative team



Weirong Chen

Jonathan Amith

Joshua Benn

Basileo Martínez  
Cruz

Haskins Labs

Gettysburg College

SUNY Buffalo

Tlaxiaco, Mexico

Not pictured: Richard Hatcher (SUNY Buffalo), Wilibaldo Martínez Cruz (Tlaxiaco, Mexico), Rey Castillo García (Guerrero, Mexico).

Special thanks: Doug Whalen (Haskins Labs/CUNY).

This material is based upon work supported by the NSF-NEH Documenting Endangered Languages Program and the NSF Robust Intelligence Program under Award No. 1603323.

## Appendix A: Methods

- Corpus of 6 speakers (3 male, 3 female) producing spontaneous narratives in YM, totalling 107 minutes; force-aligned and corrected.
- Analysis of duration and percentage of voicing (our measure of lenition) during constriction/closure for /t, k, k<sup>w</sup>, s, ʃ, h, tʃ/. A total of 7892 segments were analyzed.
- Words here were coded by stem position (initial, medial, final syllable), and word size (monosyllabic, disyllabic, polysyllabic).

- Duration was extracted with an existing Praat script and voicing was extracted with a script written for Matlab. The percentage of voicing during constriction was calculated using a normalized low frequency energy ratio (Kasi and Zahorian, 2002).
- Two separate statistical analyses were run using lmerTest (Kuznetsova et al., 2017), one with duration as the dependent variable and another with percentage of voicing as the dependent variable.
- In each model, word size, word position, and consonant were treated as fixed effects while speaker and item were treated as random effects.

## 3-way categorization

Higher accuracy found – stop vs. fricative vs. sonorant (nasal or approximant). Sonorant realizations tend to be categorized as fricatives.

**/t/ Acc: 86.4%**

Predicted	stop	96%	5%	0%
	fric	5%	90%	41%
	nas	0%	5%	59%
		stop	fric	nas
		Truth		

**/k/ Acc: 77.3%**

Predicted	stop	89%	6%	0%
	fric	11%	82%	91%
	appr	0%	12%	9%
		stop	fric	appr
		Truth		

# What is forced alignment?

An automatic method of text-speech alignment.

Recognition of the speech signal is performed using a hidden Markov model (HMM), with the search path constrained to the known sequence of phonemes.

Because a Viterbi search can yield the locations of phoneme-based states as well as the state identities, phonetic alignment can be obtained by constraining the search to the known phoneme sequence.

It is “forced alignment” because the alignment is obtained by forcing the recognition result to be the proposed phonetic sequence.

# Triqui grammar/phonology

- Final syllables are bimoraic; they may be closed with a glottal coda (/CVh, CVʔ/) or open with a long vowel (/CV:/).
- Final syllables are prominent; most of the phonological contrasts occur on them. Vowels and consonants may be reduced elsewhere.
- Tone has a high morphological load in the language, marking person, verbal aspect, and a few other distinctions.

tʃa <sup>43</sup>	'to eat (PERF)'	tʃa <sup>2</sup>	'to eat (POT)'
tʃah <sup>4</sup>	'I ate'	tʃah <sup>1</sup>	'I will eat'
tʃa <sup>41</sup> = reʔ <sup>1</sup>	'You ate'		
tʃah <sup>3</sup>	'(aforementioned) ate'	tʃah <sup>23</sup>	'(aforementioned) will eat'
tʃoʔ <sup>4</sup>	'We ate'	tʃoʔ <sup>2</sup>	'We will eat'

# Montreal Forced Aligner

In order to force align speech with MFA, one needs the following:

- 1 sound recordings with minimal sampling rate of 16 kHz
- 2 corresponding TextGrid files with identical names
- 3 Pronunciation Dictionary\*
- 4 MFA software itself (out of the box)

# What does the Corpus look like?

The content of a speech corpus can play a big role in the necessary steps in training and utilizing a forced aligner model.

Is the corpus primarily natural discourse or the results of controlled experiments?

In the former case, must decide what to with the following:

- Code Switching/Mixing
- Disfluencies

If you decide to disregard these phenomena, they must be removed from the TextGrids before running MFA.

However, it may be a good idea to keep this data, especially if your corpus is modest.

## Preparing the data - TextGrids

Beginning with ELAN annotations, we created and utilized a Python script create a new surface-true tier for each speaker in a recording.

- Remove edited insertions, i.e. annotation of either intended or unintended elided elements.
- Remove coding which identifies text as disfluencies or as Spanish
- Remove all tiers that are not the actual transcription
- Treat all non-linguistic annotation, e.g. laughing, coughing as *spn*
- Export new tiers to TextGrid file

taj13 ki3hyaj3 nni4=(reh1) yoj3 → taj13 ki3hyaj3 nni4 yoj3

be4=nih2unj4 ku3man4 \*\*sesenta\* ni2 \*\*sesenta y cinco\* bin3 →  
be4 nih2unj4 ku3man4 sesenta ni2 sesenta y cinco bin3

## Considerations for constructing an aligner

We must construct a pronunciation dictionary; a mapping between the transcription and the surface phonological shape.

Example: 'sit' SS IH1 TQ (Arpabet)

There are existing pronunciation dictionaries for well-studied languages like English and Spanish, but none for most endangered languages.

For Triqui words, Python scripts were used to create a pronunciation of each word encoded in X-SAMPA. We decided that certain rimes difficult to segment would be treated as one phone segment.

ni2	→	n i~
bin3	→	B i~
ki3hyaj	→	k i ?J aH

1. (...) marks elided speech  
 taj13 ki3hyaj3 nni4=(reh1) yoj3  
 like.so did mother=2S then  
 'Your mother did (it) like that then.'
  
2. \*\*...\* marks another language  
 be4=nih2unj4 ku3man4 \*\*sesenta\* ni2 \*\*sesenta y cinco\* bin3  
 TOP=PL.1P PERF.exist sixty and sixty five be  
 'We were (there) in (19)60 and (19)65, it was...'
  
3. [...] marks disfluencies  
 ta1ranh3 nej3 sinj5 bin3... [ranh]  
 all 3P people be ??  
 '...all of them that were there'
  
4. Loanwords use Triqui orthography  
 sa4na43 'manzana' (apple)  
 skwe4la43 'escuela' (school)

## Preparing the data - Dictionary

Although technically, MFA can run without a pronunciation dictionary, in most cases this is a crucial element of training an aligner.

The function of the pronunciation dictionary is to tell MFA what sounds to look for when encountering a particular word.

Itunyoso Triqui orthography is relatively shallow and surface-true but we decided to create a dictionary for the following reasons:

- ① Wanted MFA to disregard tone
- ② The grapheme <n> serves two functions in this orthography, the nasal stop [n], e.g. *ni<sup>2</sup>* [ni<sup>2</sup>] 'and' and to indicate that the preceding vowel is a nasal vowel, e.g. *bi<sup>3</sup>* [βi<sup>3̃</sup>] 'to be'.

## Developing the Dictionary - Triqui words

With Python scripts, we collected all the transcriptions from all the recordings in the corpus.

These were then separated into Triqui and Spanish data and both sets were then tokenized to create a word list of unique word forms, including partial words.

For Triqui words, scripts were used to create a pronunciation of each word encoded in X-SAMPA. We decided that certain rimes difficult to segment would be treated as one phone segment.

ni2	→	n i~
bin3	→	B i~
ki3hyaj	→	k i ?J aH

- Adda-Decker, M. and Snoeren, N. D. (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, 39:261–270.
- Babinski, S., Dockum, R., Goldenberg, D., Craft, J. H., Fergus, A., and Bower, C. (2019). A Robin Hood approach to Forced Alignment: English-trained algorithms and their use on Australian languages. In *Proceedings of the 93<sup>rd</sup> Annual Meeting of the Linguistic Society of America*. Linguistic Society of America.
- Bouavichith, D. and Davidson, L. (2013). Segmental and prosodic effects on intervocalic voiced stop reduction in connected speech. *Phonetica*, 70:182–206.
- Castillo García, R. (2007). Descripción fonológica, segmental, y tonal del Mixteco de Yoloxóchitl, Guerrero. Master's thesis, Centro de Investigaciones y Estudios Superiores en Antropología Social (CIESAS), México, D.F.
- Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61:30–47.
- Coto-Solano, R. A. (2017). *Tonal reduction and literacy in Me'phaa Vátháá*. PhD thesis, University of Arizona.
- Davidson, L. (2011). Characteristics of stop releases in American English spontaneous speech. *Speech Communication*, 53:1042–1058.
- Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, 54:35–50.
- Davidson, L. (2018). Phonation and laryngeal specification in American English voiceless obstruents. *Journal of the International Phonetic Association*, 48(3):331–356.
- Davis, B. L., Jakielski, K. J., and Marquardt, T. P. (1998). Developmental apraxia of speech: Determiners of differential diagnosis. *Clinical Linguistics & Phonetics*, 12(1):25–45. 

- DiCano, C. (2012a). Coarticulation between tone and glottal consonants in Itunyoso Trique. *Journal of Phonetics*, 40(1):162–176.
- DiCano, C., Amith, J. D., and Castillo García, R. (2014). The phonetics of moraic alignment in Yoloxóchitl Mixtec. In *Proceedings of the 4<sup>th</sup> Tonal Aspects of Language Symposium*. Nijmegen, the Netherlands.
- DiCano, C., Benn, J., and Castillo García, R. (2018). The phonetics of information structure in Yoloxóchitl Mixtec. *Journal of Phonetics*, 68:50–68.
- DiCano, C., Chen, W.-R., Benn, J., Amith, J. D., and Castillo García, R. (2017). Automatic detection of extreme stop allophony in Mixtec spontaneous speech. In *Annual Meeting in Phonology*. New York University.
- DiCano, C., Nam, H., Amith, J. D., Castillo García, R., and Whalen, D. H. (2015). Vowel variability in elicited versus spontaneous speech: evidence from Mixtec. *Journal of Phonetics*, 48:45–59.
- DiCano, C., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., and Castillo García, R. (2013). Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *Journal of the Acoustical Society of America*, 134(3):2235–2246.
- DiCano, C. and Whalen, D. H. (2015). The interaction of vowel length and speech style in an arapaho speech corpus. In *Proceedings of the 18<sup>th</sup> International Congress of the Phonetic Sciences*, Glasgow, Scotland.
- DiCano, C., Zhang, C., Whalen, D. H., and Castillo García, R. (2019). Phonetic structure in Yoloxóchitl Mixtec consonants. *Journal of the International Phonetic Association*, <https://doi.org/10.1017/S0025100318000294>.

- DiCano, C. T. (2008). *The Phonetics and Phonology of San Martín Itunyoso Trique*. PhD thesis, University of California, Berkeley.
- DiCano, C. T. (2010). Illustrations of the IPA: San Martín Itunyoso Trique. *Journal of the International Phonetic Association*, 40(2):227–238.
- DiCano, C. T. (2012b). Coarticulation between Tone and Glottal Consonants in Itunyoso Trique. *Journal of Phonetics*, 40:162–176.
- DiCano, C. T. (2012c). The Phonetics of Fortis and Lenis Consonants in Itunyoso Trique. *International Journal of American Linguistics*, 78(2):239–272.
- DiCano, C. T. (2016). Abstract and concrete tonal classes in Itunyoso Trique person morphology. In Palancar, E. and Léonard, J.-L., editors, *Tone and Inflection: New Facts and New Perspectives*, volume 296 of *Trends in Linguistics Studies and Monographs*, chapter 10, pages 225–266. Mouton de Gruyter.
- Evans, N., Fletcher, J., and Ross, B. (2008). Big words, small phrases: Mismatches between pause units and the polysynthetic word in Dalabon. *Journal of Linguistics*, 46(1):89–129.
- Fletcher, J. and Evans, N. (2002). An acoustic phonetic analysis of intonational prominence in two Australian languages. *Journal of the International Phonetic Association*, 32:123–140.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., and Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.
- Hualde, J. I., Simonet, M., and Nadeu, M. (2011). Consonant lenition and phonological recategorization. *Journal of Laboratory Phonology*, 2:301–329.
- Johnson, L. M., DiPaolo, M., and Bell, A. (2018). Forced Alignment for Understudied Language Varieties: Testing ProsodyLab-Aligner with Tongan Data. *Language Documentation and Conservation*, 12:80–123.

- Kakadelis, S. M. (2018). *Phonetic Properties of Oral Stops in Three Languages with No Voicing Distinction*. PhD thesis, Graduate Center, City University of New York.
- Kaland, C. and Himmelmann, N. P. (2019). Repetition reduction revisited: The Prosody of Repeated Words in Papuan Malay. *Language and Speech*, pages 1–25.
- Kasi, K. and Zahorian, S. A. (2002). Yet another algorithm for pitch tracking. In *Proceedings of ICASSP02*, pages 361–364. Orlando.
- Katz, J. (2016). Lenition, perception, and neutralisation. *Phonology*, 33:43–85.
- Katz, J. and Fricke, M. (2018). Auditory disruption improves word segmentation: A functional basis for lenition phenomena. *Glossa*, 3(1):1–25.
- Katz, J. and Pitzanti, G. (2019). The phonetics and phonology of lenition: A Campidanese Sardinian case study. *Journal of Laboratory Phonology*, 10(1):1–40.
- Kirschenbaum, A., Wittenburg, P., and Heyer, G. (2012). Unsupervised morphological analysis of small corpora: First experiments with Kilivila. In Seifart, F., Haig, G., Himmelmann, N. P., Jung, D., Margetts, A., and Trilsbeek, P., editors, *Potentials of Language Documentation: Methods, Analyses, and Utilization*, chapter 4, pages 25–31. Language Documentation & Conservation Special Publication No. 3.
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26.
- Ladd, D. R. (2014). *Simultaneous Structure in Phonology*. Oxford University Press.
- Lewis, A. M. (2001). *Weakening of intervocalic /ptk/ in two Spanish dialects: Toward the quantification of lenition processes*. PhD thesis, University of Illinois at Urbana-Champaign.
- Lin, C.-Y., Roger Jang, J.-S., and Chen, K.-T. (2005). Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS. *Computational Linguistics and Chinese Language Processing*, 10(2):145–166.

- Lin, J., Li, W., Gao, Y., Xie, Y., Chen, N. F., Siniscalchi, S. M., Zhang, J., and Lee, C.-H. (2018). Improving Mandarin Tone Recognition Based on DNN by Combining Acoustic and Articulatory Features Using Extended Recognition Networks. *Journal of Signal Processing Systems*, 90:1077–1087.
- Malfrère, F., Deroo, O., Dutoit, T., and Ris, C. (2003). Phonetic alignment: speech synthesis-based vs. Viterbi-based. *Speech Communication*, 40:503–515.
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings from Interspeech 2017*.
- Michaud, A., Adams, O., Cohn, T. A., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone Toolkit. *Language Documentation and Conservation*, 12:393–429.
- Mitra, V., Kathol, A., Amith, J. D., and Castillo García, R. (2016). Automatic speech transcription for low-resource languages - The Case of Yoloxóchtli Mixtec (Mexico). In *Interspeech 2016*. ISCA, San Francisco.
- Muehlbauer, J. (2012). Vowel spaces in Plains Cree. *International Journal of American Linguistics*, 42(1):91–105.
- Ní Chasaide, A., Wogan, J., Ó Raghallaigh, B., Ní Bhriain, A., Zoerner, E., Berthelsen, H., and Gobl, C. (2006). Speech technology for minority languages: the case of Irish (Gaelic). In *INTERSPEECH-2006*, pages 181–184.
- Palancar, E. L., Amith, J. D., and Castillo García, R. (2016). Verbal inflection in Yoloxóchtli Mixtec. In Palancar, E. L. and Léonard, J.-L., editors, *Tone and Inflection: New Facts and New Perspectives*, chapter 12, pages 295–336. Mouton de Gruyter. 

- Roux, J. C. and Visagie, A. S. (2007). Data-driven approach to rapid prototyping Xhosa speech synthesis. In *Proceedings of the 6<sup>th</sup> ISCA Workshop on Speech Synthesis*, pages 143–147.
- Schreer, O. and Schneider, D. (2012). Supporting linguistic research using generic automatic audio/video analysis. In Seifart, F., Haig, G., Himmelmann, N. P., Jung, D., Margetts, A., and Trilsbeek, P., editors, *Potentials of Language Documentation: Methods, Analyses, and Utilization*, chapter 6, pages 39–45. Language Documentation & Conservation Special Publication No. 3.
- Stuart-Smith, J., Sonderegger, M., Rathcke, T., and Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, 6(3-4):505–549.
- Tang, K. and Bennett, R. (2018). Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan). *Journal of the Acoustical Society of America*, 144(2):997–1017.
- Tang, K. and Bennett, R. (2019). Unite and conquer: bootstrapping forced alignment tools for closely-related minority languages Mayan. In Calhoun, S., Escudero, P., Tabain, M., and Warren, P., editors, *Proceedings of the International Congress of Phonetic Sciences (ICPhS) 2019*, pages 3584–3552. Canberra, Australia: ASSTA.
- Torreira, F. and Ernestus, M. (2011). Realization of voiceless stops and vowels in conversational French and Spanish. *Journal of Laboratory Phonology*, 2:331–353.
- Warner, N. and Tucker, B. V. (2011). Phonetic variability of stops and flaps in spontaneous and careful speech. *Journal of the Acoustical Society of America*, 130(3):1606–1617.
- Wedel, A., Kaplan, A., and Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2):179–186.

- Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics - 2008*.
- Yuan, J. and Liberman, M. (2009). Investigating /l/ variation in English through forced alignment. In *Interspeech - 2009*, pages 2215–2218.