Constructing a forced alignment system for Itunyoso Triqui: Challenges, Outcomes, and Opportunities

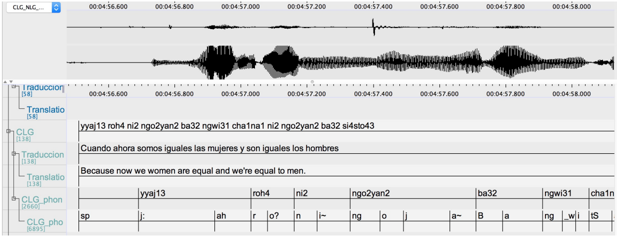
Richard Hatcher
rjhatche@buffalo.edu
Christian DiCanio
cdicanio@buffalo.edu

Department of Linguistics
University at Buffalo

9/27/19

The problem: fieldwork → spoken language corpus

↓



yyaj13 roh4 ni2 ngo2yan2 ba32 ngwi31 cha1na1 ni2 ngo2yan2 ba32 si4sto43

Cuando ahora somos iguales las mujeres y son iguales los hombres

Because now we women are equal and we're equal to men.

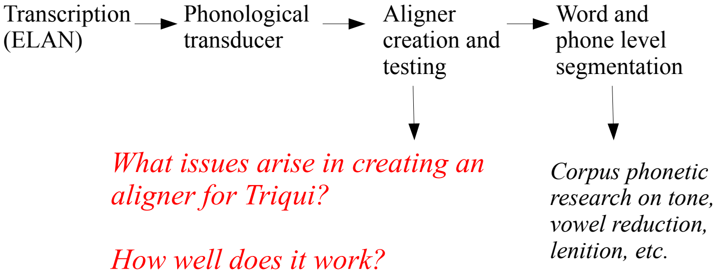
yyaj13 roh4 ni2 ngo2yan2 ba32 ngwi31 cha1n

sp j: ah r o? n i~ ng o j a~ B a ng _w i tS

The documentation → speech corpus framework

The typical framework for language documentation involves audio/video recording, linguistic description, and transcription of about 30-40 hours of speech.

Producing this is an immense labor and time commitment. Moreover, in terms of speech *corpus* development, it's at an early stage still!



Transcription (ELAN) → Phonological transducer → Aligner creation and testing → Word and phone level segmentation

Aligner creation and testing ↓ *What issues arise in creating an aligner for Triqui?*

How well does it work?

Word and phone level segmentation ↓ *Corpus phonetic research on tone, vowel reduction, lenition, etc.*

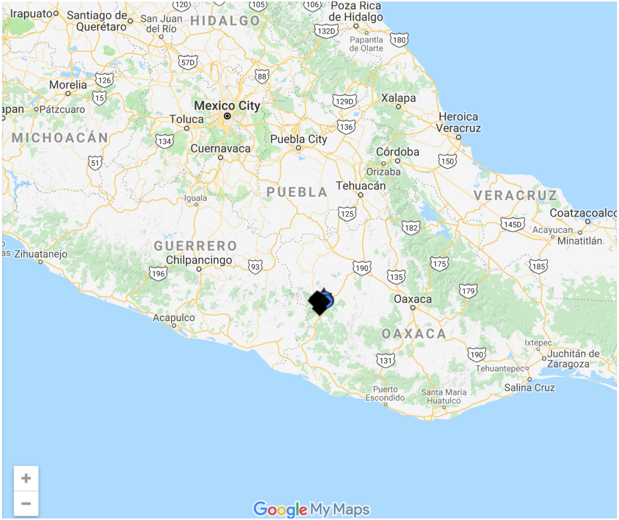
Roadmap

1. The Itunyoso Triqui corpus
2. The Whats and Whys of forced alignment
3. Tutorial on creating an aligner for a multi-lingual documentation corpus

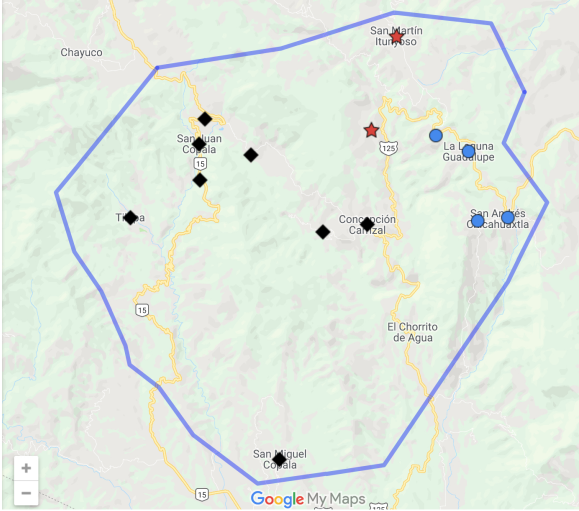
The Itunyoso Triqui Corpus

- Otomanguean, spoken in Oaxaca, Mexico (∼2500 speakers).
- Running speech: 25 hours of transcribed personal narratives, stories, and folklore; 31 speakers. (Another 5-10 hours are untranscribed).
- Collected most narratives between 2013 - 2017.
- Initial transcription done by trained native speakers, subsequent revision with PI (DiCanio).
- Experimental work: Approximately 50 hours of recordings; from 2004 - present.
- Archived at AILLA (Archive of the Indigenous Languages of Latin America)
- Phonological/phonetic fieldwork (DiCanio, 2008, 2010, 2012b,a, 2016)

Triqui region

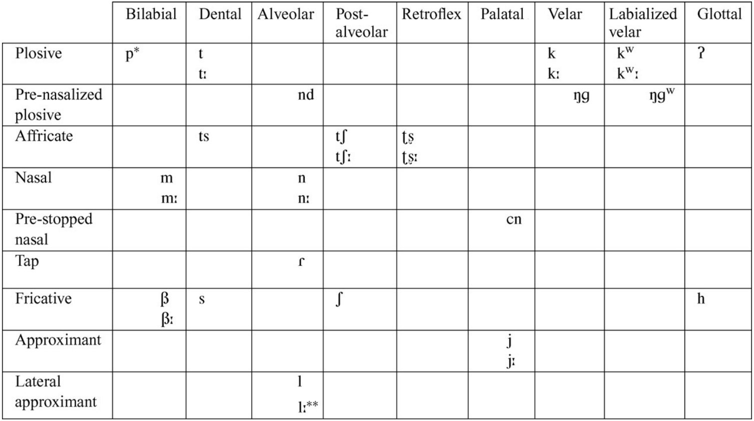


Triqui region: 3 dialects (colors)



Segmental inventory

(DiCanio, 2010)



	Bilabial	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Labialized velar	Glottal
Plosive	p*	t tː					k kː	kʷ kʷː	ʔ
Pre-nasalized plosive			nd				ŋg	ŋgʷ	
Affricate		ts		tʃ tʃː	ʈʂ ʈʂː				
Nasal	m mː		n nː						
Pre-stopped nasal						cn			
Tap			ɾ						
Fricative	β βː	s		ʃ					h
Approximant						j jː			
Lateral approximant			l lː**						

*Rare in native words **Occurs in only a few words

Glottalized consonants

	Bilabial	Alveolar	Palatal	Velar
Pre-nasalized plosive		ˀnd		ˀŋg
Nasal	ˀm	ˀn		
Trill		ˀr̥*		
Fricative	ˀβ			
Approximant			ˀj	
Lateral approximant		ˀl		

*Occurs in one lexical item

Vowels

	Front		Central		Back	
	Oral	Nasal	Oral	Nasal	Oral	Nasal
Close	i	ĩ			u	ũ
Close-mid	e			ə̃	o	
Open			a			

Tone
(DiCanio, 2016)

Nine contrastive tones on root-final syllables; fewer on non-final syllables, prefixes, and clitics.

	Open syllable		Coda /h/		Coda /ʔ/	
Tone	Word	Gloss	Word	Gloss	Word	Gloss
/4/	yũ^{4}	‘earthquake’	yãh^{4}	‘dirt’	niʔ^{4}	‘see.1DU’
/3/	yũ^{3}	‘palm leaf’	yãh^{3}	‘paper’	tsiʔ^{3}	‘pulque’
/2/	ũ^{2}	‘nine’	tah^{2}	‘delicious’	ttʃiʔ^{2}	‘ten’
/1/	yũ^{1}	‘loose’	kãh^{1}	‘naked’	tsiʔ^{1}	‘sweet’
/45/			toh^{45}	‘forehead’		
/13/	yo^{13}	‘fast (adj.)’	toh^{13}	‘a little’		
/43/	ɾa^{43}	‘want’	nnãh^{43}	‘mother!’		
/32/	ɾã^{32}	‘durable’	nnãh^{32}	‘cigarette’		
/31/	ɾã^{31}	‘lightning’				

Triqui grammar/phonology

- Final syllables are bimoraic; they may be closed with a glottal coda (/CVh, CVʔ/) or open with a long vowel (/CVː/).
- 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^{43}	'to eat (PERF)'	tʃa^{2}	'to eat (POT)'
tʃah^{4}	'I ate'	tʃah^{1}	'I will eat'
tʃa^{41}=ɾeʔ^{1}	'You ate'		
tʃah^{3}	'(aforementioned) ate'	tʃah^{23}	'(aforementioned) will eat'
tʃoʔ^{4}	'We ate'	tʃoʔ^{2}	'We will eat'

The transcriptional orthography varies a little bit from the phonological representation, but all contrasts are maintained. Tone is marked after each syllable.

nũ^3	ki^{1}ɾiʔ^1	ʈʂa^3	tʃoʔ^2	ɾah^4	IPA
nun3	ki1-rih1	chra3	choh2	raj4	Practical transcription
NEG	POT-get	tortilla	eat.POT.1P	believe	Gloss
'We couldn't find (any) tortillas for us to eat, I think.'					Translation

tːũh^2	tu^{3}k^{w}a^{3}tʃiʔ^3	a^{3}ʔnĩh^5 = neh^3	ɾĩãh^3	nã^2 ju^{3}βe^{32}
ttunj2	tu3kwa3chih3	a3hninj5 = nej3	rianj3	nan2 yu3be32
eight	pair.of.thread	insert = 3P	in.3PS	then
'Eight pairs of threads they put in it then.'				

Why segment the Triqui corpus?

1. Acoustic segmentation of speech data is the preliminary step to extracting phonetic data from the speech signal for corpus phonetics.

2. Segmented speech is useful for localization of words and morphological boundaries. This is relevant for dictionary work, usage-based linguistic analysis, discourse analysis, and other areas.

3. Future annotation of the speech signal requires initial segmentation.

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.

- Forced aligners are language-specific. They are often trained on a corpus of data from a language where segmentation by hand has already been done. These data are used to build hidden markov models for the acoustic signature of each phone. The forced alignment system then uses its internal model to predict where boundaries between phones occur.

- Most systems are trained on major languages such as English (Malfrère et al., 2003; Yuan and Liberman, 2008, 2009), French (Adda-Decker and Snoeren, 2011; Malfrère et al., 2003), Spanish (Malfrère et al., 2003), Dutch (Malfrère et al., 2003), and Mandarin Chinese (Lin et al., 2005) but a few are trained on less well-studied languages like Gaelic (Ní Chasaide et al., 2006) and Xhosa (Roux and Visagie, 2007).

Assessing alignment

Existing aligners can be used to align a corpus speech for which no aligner exists (c.f. DiCanio et al. (2013)), but accuracy is not as good as using an aligners trained on the target language.

Alignment is assessed by determining how far off acoustic boundaries are between automatic segmentation and *human* segmentation. Example below using the Montreal Forced Aligner for English (McAuliffe et al., 2017).

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

Assessing the Triqui aligner

We trained a Triqui aligner on approximately 5.5 hours of transcribed Triqui texts (running speech); 88 sound files.

We then selected four texts totalling 33.8 minutes (7 speakers) for which we had human labelling (gold standard).

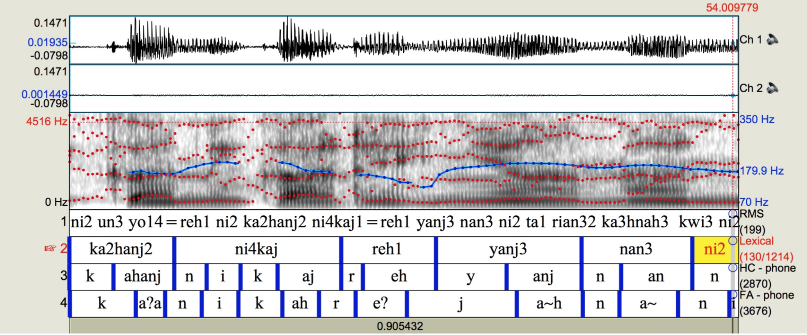
We compared forced alignment with the Triqui aligner against the human labeller.

While this may not seem like much speech to examine - it's 16,553 speech segments!

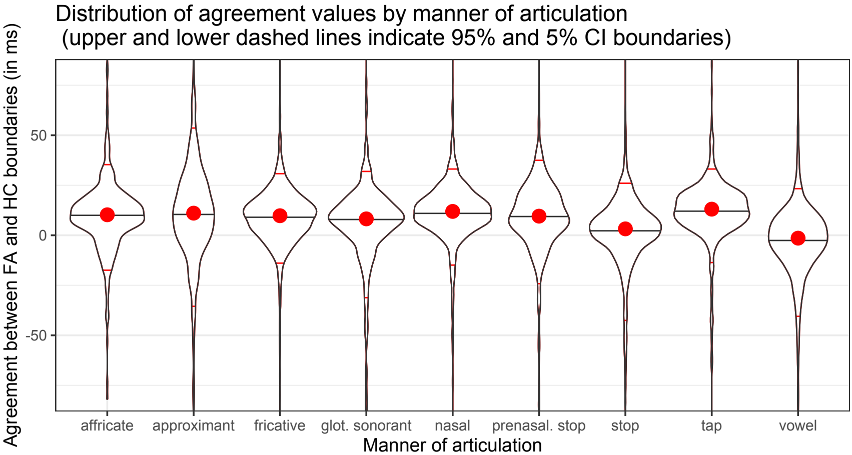
Results

Our alignment was quite good compared to the MFA system for English (McAuliffe et al., 2017).

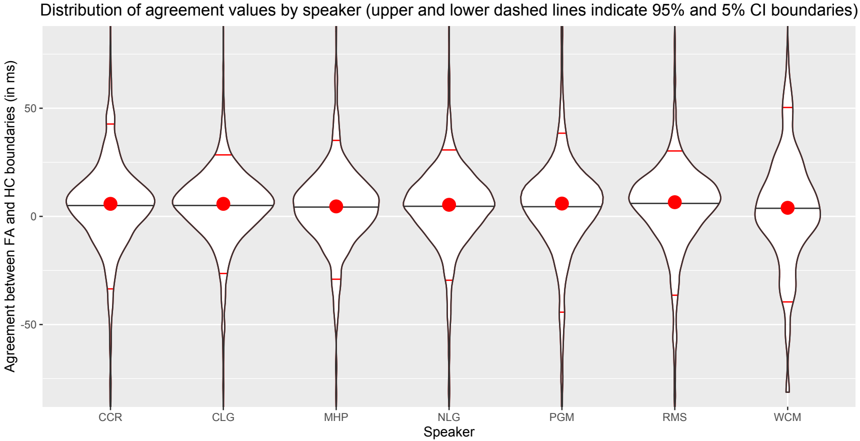
Tolerance	10 ms	20 ms	30 ms	40 ms	50 ms
% phones in corpus	46.7%	77.1%	89.2%	93.7%	95.9%


54.009779
0.1471
0.01935
-0.0798
0.1471
0.001449
-0.0798
Ch 1
Ch 2
4516 Hz
0 Hz
350 Hz
179.9 Hz
70 Hz
1 ni2 un3 yo14=reh1 ni2 ka2hanj2 ni4kaj1=reh1 yanj3 nan3 ni2 ta1 rian32 ka3hnah3 kwi3 ni2
RMS (199)
2 ka2hanj2 ni4kaj reh1 yanj3 nan3 ni2
Lexical (130/1214)
3 k ahanj n i k aj r eh y anj n an n
HC - phone (2870)
4 k a?a n i k ah r e? j a~h n a~ n i
FA - phone (3676)
0.905432

Alignment for affricates was significantly better than other manners of articulation (low variance); alignment for stops and approximants was worse (higher variance).



Alignment did not differ much by speaker, but in general it was slightly better for female speakers than male speakers.



Preliminary considerations for constructing an aligner

For the Triqui corpus, we have recordings and transcriptions (in ELAN, but exportable to many formats).

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.

Pre-preliminaries (ATTN: fieldworkers!)

If your orthographic transcription of the language is fairly phonemic and reflects surface phonological structure, creating a set of pronunciations for all words in the corpus is fairly straightforward. There are additional issues that may arise though:

- Many endangered language corpora are multilingual; how do you separate different languages?
- How are loanwords tagged?
- How are disfluencies tagged?
- How are elided tokens tagged?
- (MOST important) What is your level of transcription? (morphophonological, phonemic, surface phonological, phonetic)

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)

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

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*^{2} [ni^2] ‘and’ and to indicate that the preceding vowel is a nasal vowel, e.g. *bin*^{3} [βĩ^3] ‘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

Developing the Dictionary - Spanish words

For Spanish words, we have access to an existing Spanish pronunciation dictionary.

We collected the pronunciation of each Spanish word in our corpus from this dictionary.

If the word was not found in the dictionary, we simply made a pronunciation for it.

The Spanish words and pronunciations were added to the Triqui dictionary.

abrieron	→	a B r j e r o n
Chicahuaxtla	→	C i k a w a s t l a

DEMO!

Two demonstrations:

- Train a model on a very VERY small corpus.
- Use: bin/mfa_train_and_align input dir. dictionary output dir.
- Align same recording with a pre-trained model.
- Use: bin/mfa_align input dir. dictionary model output dir.

Adda-Decker, M. and Snoeren, N. D. (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, 39:261–270.

DiCanio, 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.

DiCanio, C. T. (2008). *The Phonetics and Phonology of San Martín Itunyoso Trique*. PhD thesis, University of California, Berkeley.

DiCanio, C. T. (2010). Illustrations of the IPA: San Martín Itunyoso Trique. *Journal of the International Phonetic Association*, 40(2):227–238.

DiCanio, C. T. (2012a). Coarticulation between Tone and Glottal Consonants in Itunyoso Trique. *Journal of Phonetics*, 40:162–176.

DiCanio, C. T. (2012b). The Phonetics of Fortis and Lenis Consonants in Itunyoso Trique. *International Journal of American Linguistics*, 78(2):239–272.

DiCanio, 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.

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.

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*.

Ní Chasaide, A., Wogan, J., Ó Raghaillaigh, 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.

Roux, J. C. and Visagie, A. S. (2007). Data-driven approach to rapid prototyping Xhosa speech synthesis. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, pages 143–147.

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.