

# Uneven success: Automatic speech recognition and ethnicity-related dialects

Alicia Beckford Wassink

Department of Linguistics, University of Washington

<https://depts.washington.edu/sociolab/>

Panel Title:

Ethical risks of voice technology: A sociolinguistic perspective

14 February 2020



# Outline

I. Research Aims

II. Background

a) What's "sociophonetics?"

b) Our tool: CLOx

III. Methods

a) The sample: 4 ethnic groups from Pacific Northwest English (PNWE) study corpus

b) Targeted linguistic variables

IV. By-ethnicity results

V. Some surprising findings

VI. Conclusions

# acknowledgements

## CLOx Team:



Champion Fellin



David Nichols



Robert Squizzero

## PNWE Team:



Isabel Bartholomew



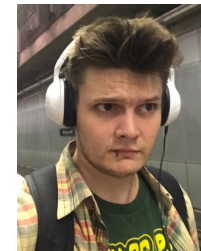
Sophia Chan



Cady Gansen



Monica Jensen



Nathan Johnson

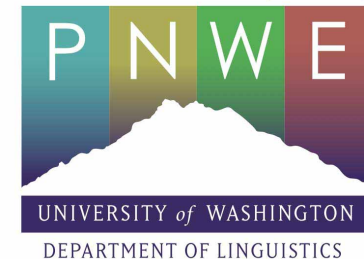


Michael Scanlon



National Science Foundation  
BCS-1844350

The Pacific Northwest English Study



# I. Aims

- Not all features of speech are handled well
- Contemporary use cases:
  - Siri, Alexa, Cortana
  - Payment-by-phone
- Inequity in access to services
- Research Questions: What differences do we observe in error types? What dialect features appear to be most challenging for our CLOx system?

## II. Background

# Sociophonetics

- A subfield of linguistics that identifies and explains socially-structured variation in the sound systems of human languages.
- Concerned with how such variation is learned, stored cognitively, subjectively evaluated, and processed in speaking and listening.

Foulkes, Scobbie and Watt 2010; diPaolo and Yaeger-Dror 2011

# Linguistic variable

- Def.: “a linguistic form whose occurrence cannot be explained without taking social characteristics into account”
- Ex. “They were the tawatees.”



Lexical variable

“local doctor, medicine person”

[deɪ wə di tawətɪz]



Phonetic variable

International Phonetic Alphabet (IPA)

(th)-stopping

Yakama English (WA)

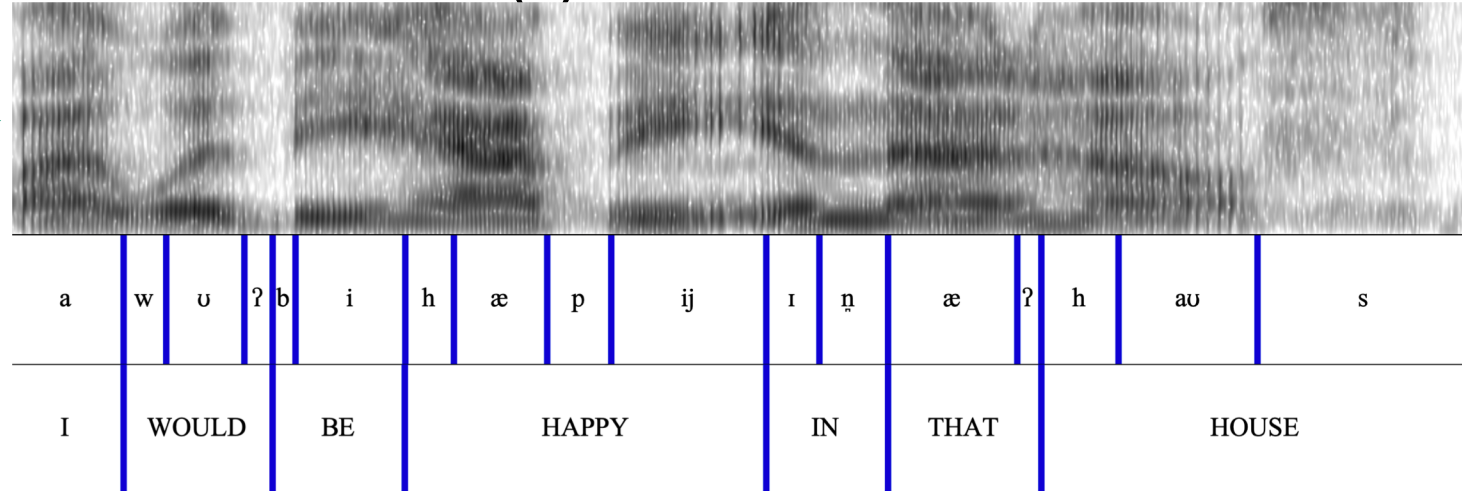
# Reading Passage example

Vowels:

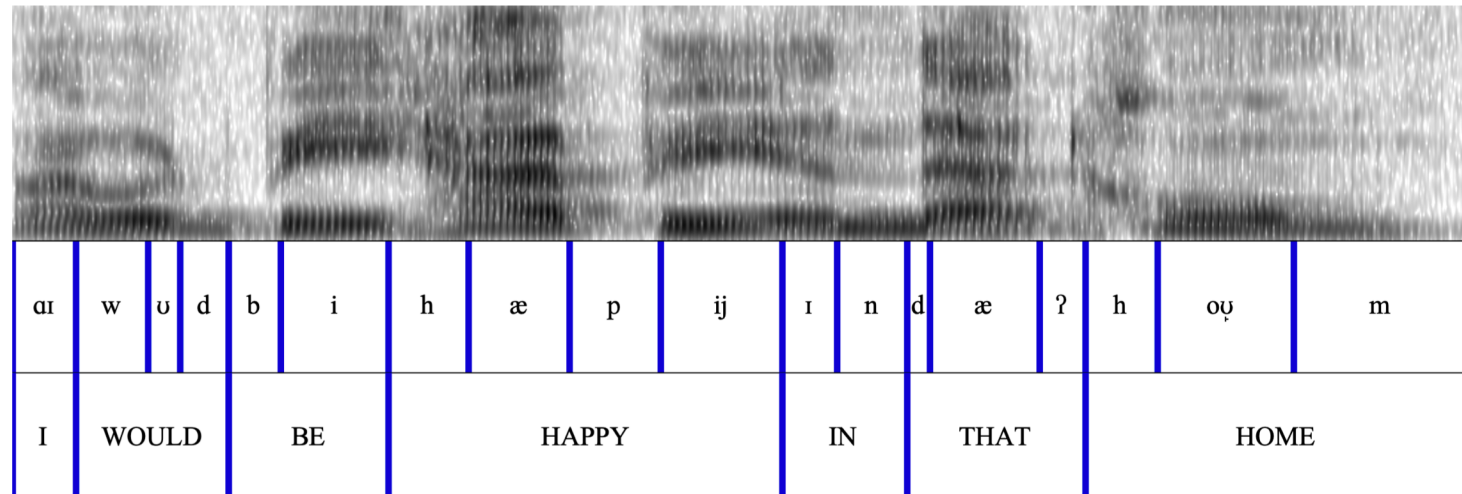
A **formant** is a concentration of acoustic energy around a particular frequency in the speech signal.



## African American (F)



## Yakama (M)





# III. Methods

# Speaker sample: 4 WA dialects



Note: Speaker classification into ethnic groups was based upon:

- Speaker's self-identification
- Social network data (membership in a speech community)
- Length of time in speech community

\* Neither dialect nor ethnic affiliation may be definitively ascertained by visual appearance.

# Tasks

- 16 speakers, 4 Ethnic groups
- Three tasks:
  - Dyadic conversations (casual, most dialectal forms)
  - Reading Passage (read, common forms)
  - Word Game task (unscripted, common forms)
- Data amounts:
  - Approx. 45 - 90 min. of speech per recording
  - 6,654 - 16,276 words per ethnic group
- Submitted to ASR tool
- Coding:
  - Manual coding in Praat (acoustic analysis software).  
Auditory analysis supplemented by use of waveform and spectrogram

# Our Tool: CLOx



- Client Libraries Oxford
- Automated audio transcription service for linguists developed by the Sociolinguistics Laboratory at the University of Washington.
- Built on the Microsoft Speech Service (via Azure subscription to Cognitive Services).
- Automatic speech recognition uses the Speech-to-text service SDK.
- CLOx delivers a conversational recording to MS Speech, which returns plain-text transcribed output, then CLOx performs output checking and supplies timestamps indicating the start and end time of each run of speech.
- We estimate that CLOx transcription is at least **five times faster** than manual transcription (hence, the logo!)

# Our Tool: CLOx



1 API KEY

2 REGION

3 LANGUAGE

4 OUTPUT FILE NAME

5 PREPROCESSING  Audio is preprocessed

Click the "Select Files and Start" button below to select audio and begin transcription. To select multiple files, use ctrl+click, cmd+click or shift+click in the file selection menu that appears after clicking.

Select Files and Start

Stop

RESULTS

Questions? Email cloxhelp at uw.edu  
Developed and maintained by the [University of Washington Sociolinguistics Laboratory](#).  
Powered by Microsoft Cognitive Services. ©2019.

# General error types

Code	Label	Example error	Target	IPA
R	reduction	lotta	lot of	varies
D	disfluencies	enough	and uh	
NC	no code	changing	digging	
NULL	words inserted	could ("windows <u>could</u> they would")	∅	
PN	Proper name	topless	Toppenish	
H	Homophone	are~R~our	are~R~our	

- Not associated with any specific dialect
- Not targeted for sociophonetic study

# Sociolinguistic Variables

## Consonants:

Code	Sociolinguistic Label	Example error	Target	IPA
(ing)	-ing (unstressed)	pick into	picking too	[ɪŋ] vs [ɪn] vs [in]
(TH)	th-stopping	den	then	/ð/ → [d]
(ʔ)	word-medial glottalization	right are	writer	/t/ → [ʔ]
(ɹ)	coda-r deletion	what a	water	/ɹ/ → ∅
(d)	consonant cluster deletion	pace [peɪs]	paced /peɪst/	/st/ → [s]
(l)	lenition	sheep	cheap	/tʃ/ → [ʃ]

- ARE associated with specific dialects
- ARE targeted for sociophonetic study



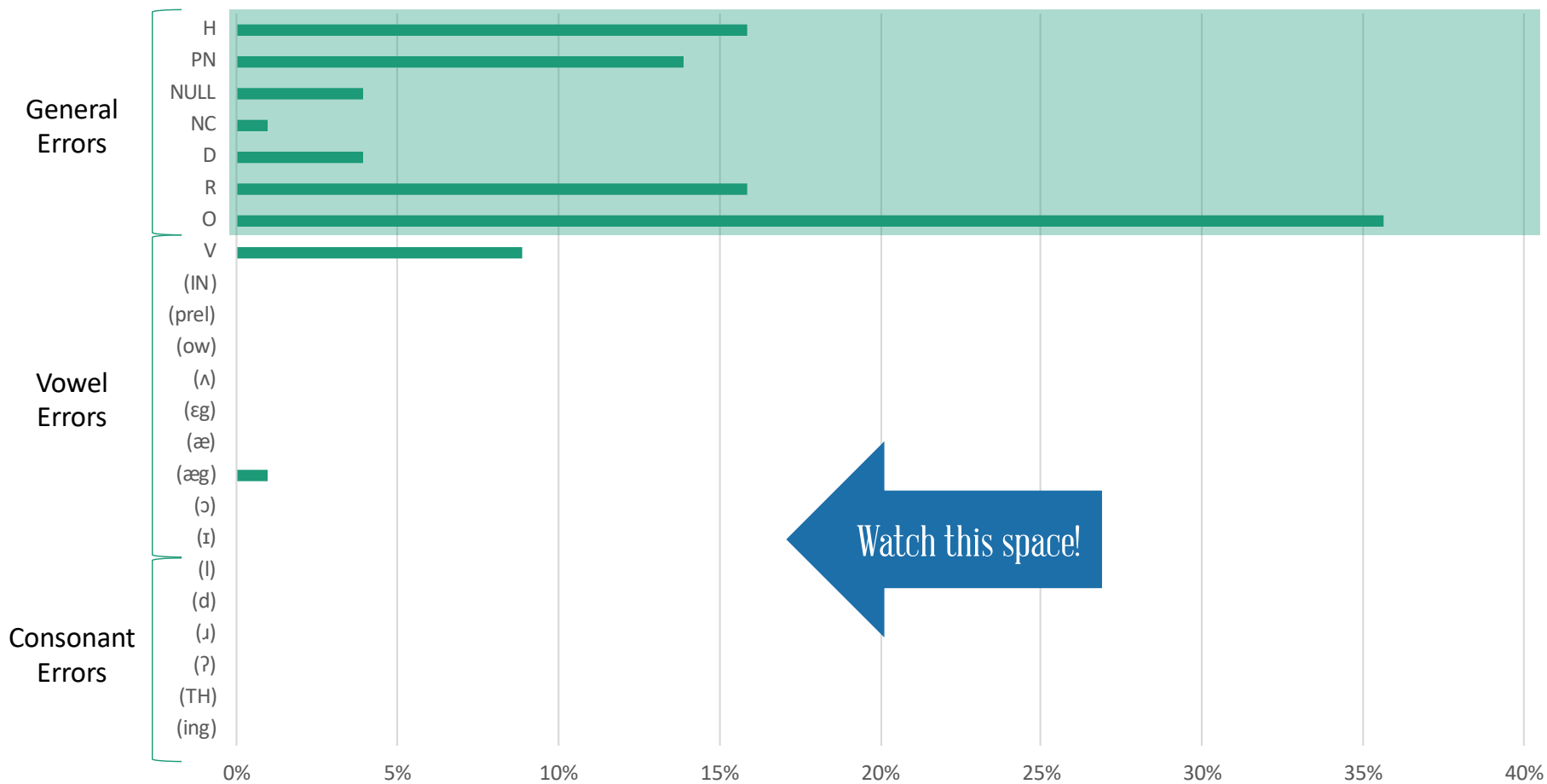
# Sociolinguistic Variables

## Vowels:

Code	Sociolinguistic Label	Example error	Target	IPA
(ɪ)	(ɪ)-tensing	peaking	picking	/ɪ/ → [i]
(ɔ)	caught/cot merger	com, cot	calm, caught	/ɔ/ → [a], /ɔ/ → [ɑ]
(æɪ)	pre-voiced velar (æ)-raising	beg	bag	/æɪ/ → [e:g]
(æ)	mistaking (æ) for other Vowel	infect	in fact	/æ/ → [a], /æ/ → [ɛ]
(ɛɪ)	pre-voiced velar (ɛ)-raising	beg	bake	/ɛɪ/ → [e:g]
(ʌ)	(ʌ)-raising	is	us	/ʌ/ → [i], /ʌ/ → [ɪ]
(ow)	(ow)-fronting	boot	boat	/ow/ → [u]
(prel)	prelateral back vowel merger	full, hole	fool, hull	/ul/ ↔ /ol/, /ʊl/ ↔ /ul/, /ʌl/ ↔ /ol/
(ɪn)	pin/pen merger	pin	pen	/ɪn/ ↔ /ɛn/
V	other vowel error	greet	great	varies
O	other (phonetic/phonological errors)	thing, faults	vague, false	varies

- ARE associated with specific dialects
- ARE targeted for sociophonetic study

## CLOx Errors, by type (Caucasian American Subsample)



Watch this space!

	(ing)	(TH)	(ʔ)	(ɹ)	(d)	(l)	(ɪ)	(ɔ)	(æg)	(æ)	(ɛg)	(ʌ)	(ow)	(prel)	(IN)	V	O	R	D	NC	NULL	PN	H
■ %	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	9%	36%	16%	4%	1%	4%	14%	16%

Example

# Normalized Frequency (*nf*)

*E* Erroneous forms across all targeted linguistic variables in a corpus

*N* Total word count for the corpus

*B* Base of normalization = 100 words

*nf*  $(E/N) * B$   
Number of error in corpus / total corpus x  
base of normalization

$$E = 668$$

$$N = 16,276$$

$$nf = (668/16276) * 100 \\ = 4.104$$

# IV. Results

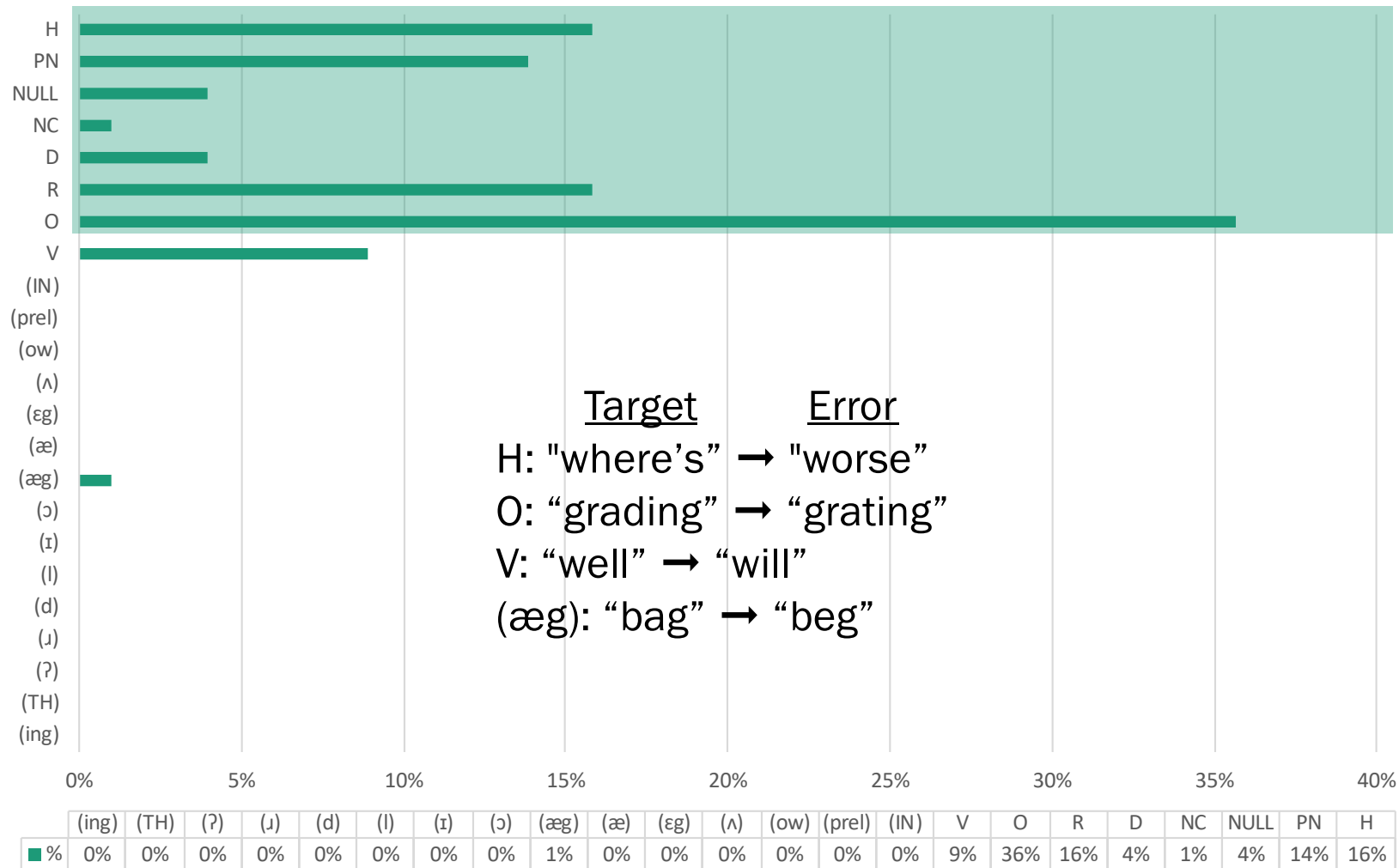
- Overall *nf*, by ethnicity

Group	N=	<i>nf</i>
Caucasian American	6,654	1.5
African American	16,276	4.1
Chicanx	3,986	8.8
Yakama	14,581	8.9

# #1: Fewest errors (nf=1.5)



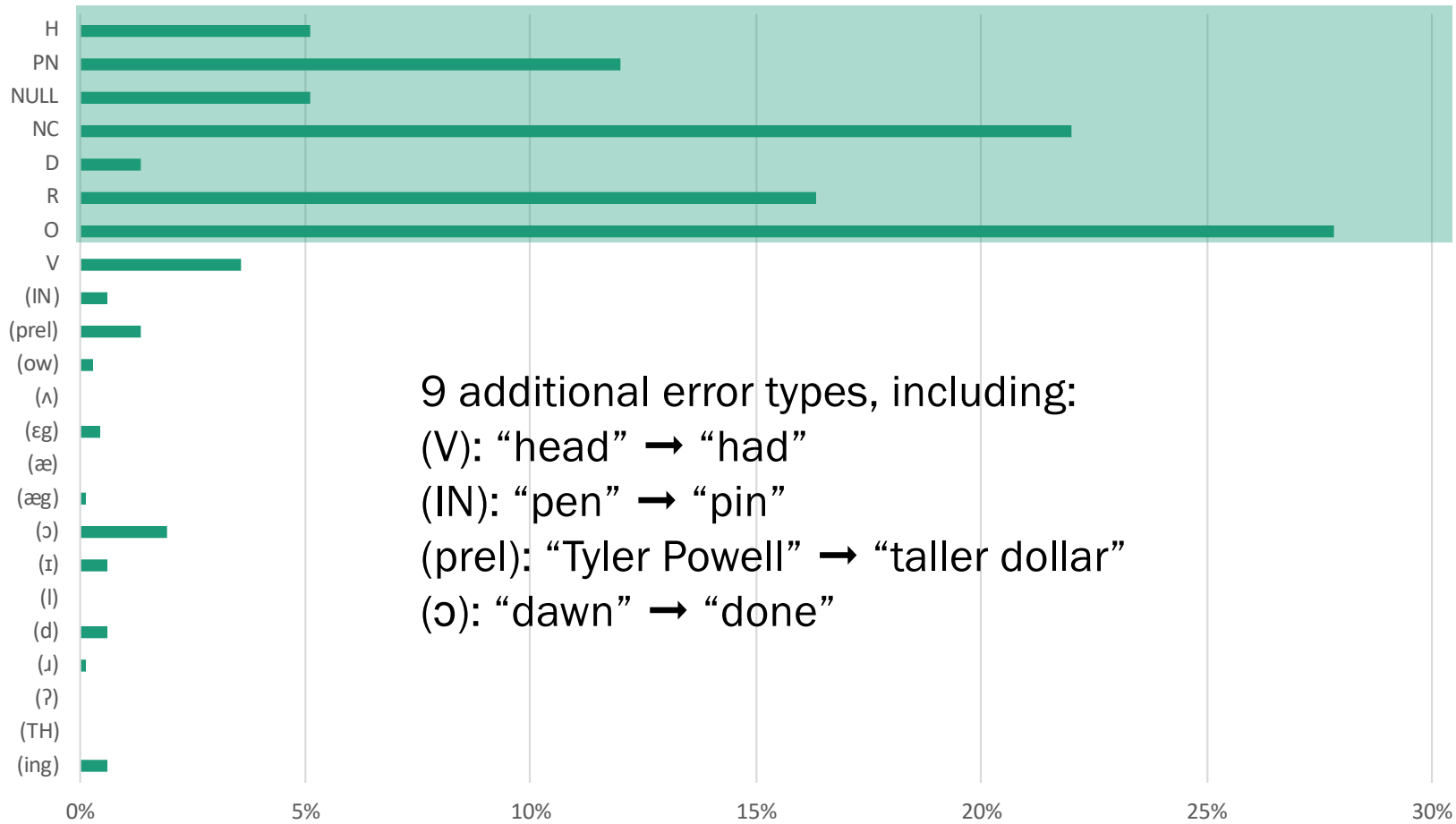
CLOx Errors, by type (Caucasian American Subsample)



## #2: (nf=4.1)



### CLOx Errors, by type (African American Subsample)



9 additional error types, including:

(V): “head” → “had”

(IN): “pen” → “pin”

(prel): “Tyler Powell” → “taller dollar”

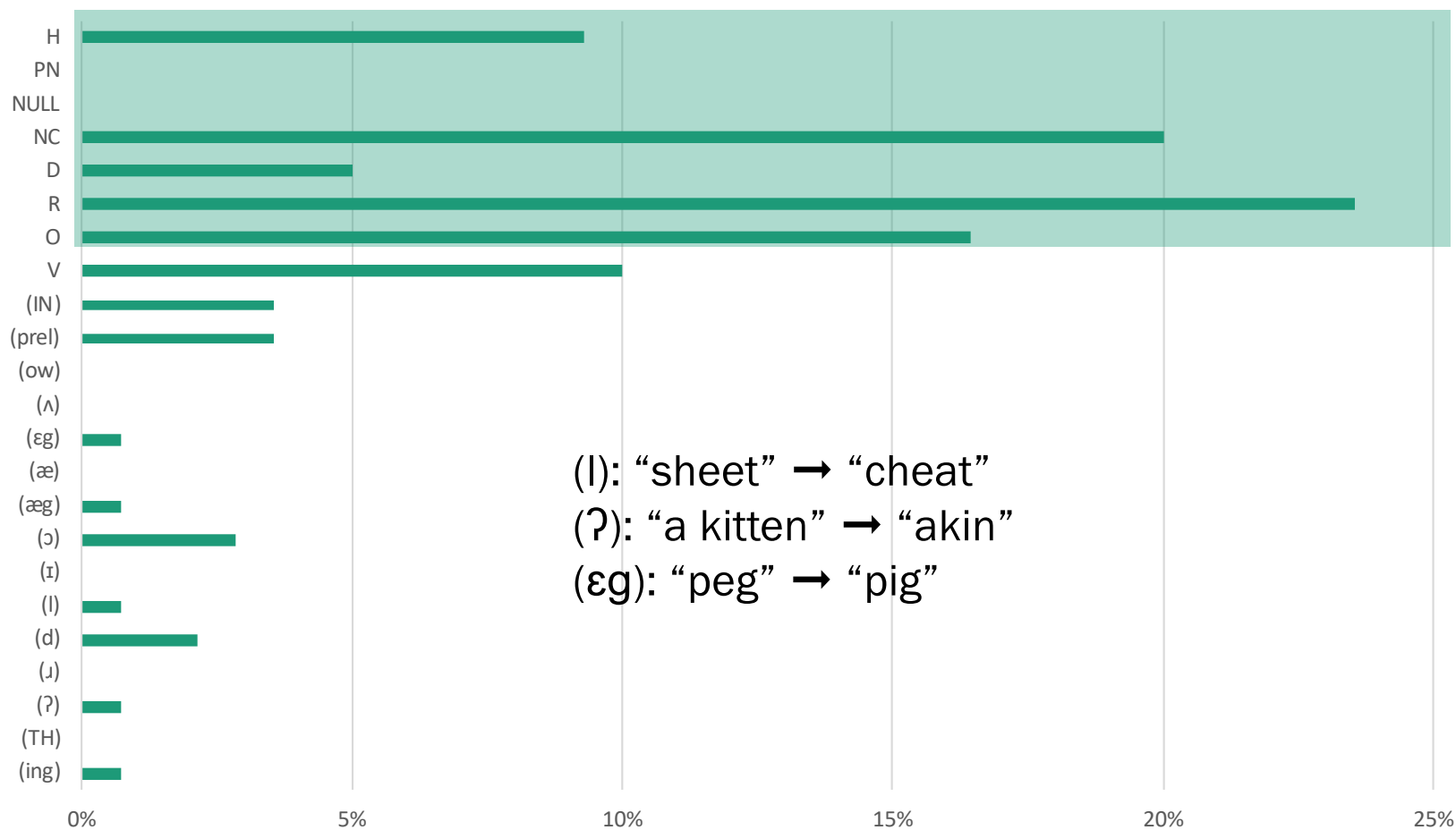
(ɔ): “dawn” → “done”

	(ing)	(TH)	(?)	(ɹ)	(d)	(l)	(ɪ)	(ɔ)	(æɟ)	(æ)	(ɛg)	(ʌ)	(ow)	(prel)	(IN)	V	O	R	D	NC	NULL	PN	H
■ %	1%	0%	0%	0%	1%	0%	1%	2%	0%	0%	0%	0%	0%	1%	1%	4%	28%	16%	1%	22%	5%	12%	5%

#3: (nf=8.8)



CLOx Errors, by type (Chicanx Subsample)



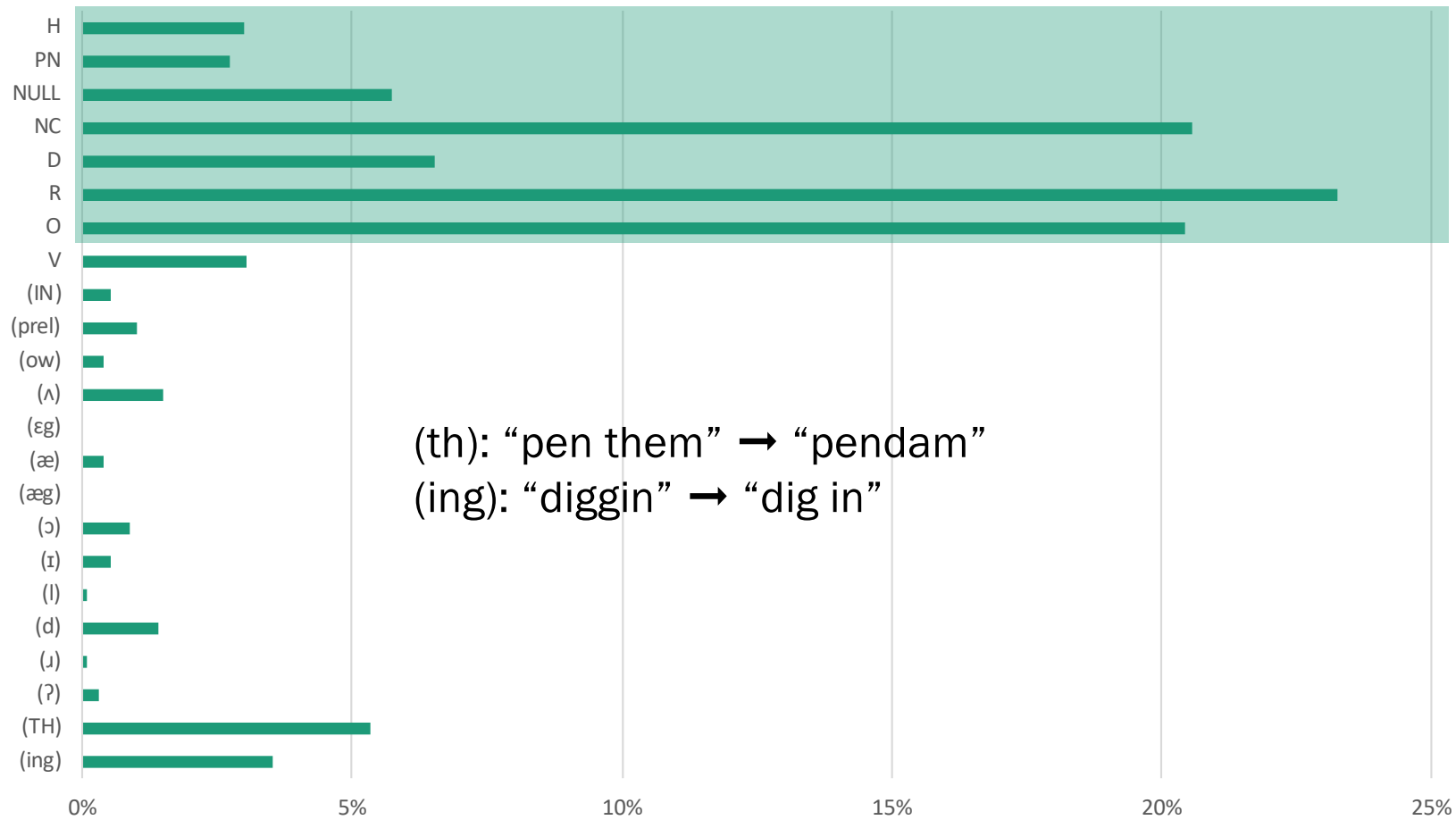
(l): “sheet” → “cheat”  
 (?): “a kitten” → “akin”  
 (ɛg): “peg” → “pig”

	(ing)	(TH)	(?)	(ɹ)	(d)	(l)	(ɪ)	(ɔ)	(æɣ)	(æ)	(ɛg)	(ʌ)	(ow)	(prel)	(IN)	V	O	R	D	NC	NULL	PN	H
■ %	1%	0%	1%	0%	2%	1%	0%	3%	1%	0%	1%	0%	0%	4%	4%	10%	16%	24%	5%	20%	0%	0%	9%

# #4: Most errors (nf=8.9)



CLOx Errors, by type (Yakama Subsample)



(th): “pen them” → “pendam”  
 (ing): “diggin” → “dig in”

	(ing)	(TH)	(ʔ)	(ɹ)	(d)	(l)	(ɪ)	(ɔ)	(æg)	(æ)	(ɛg)	(ʌ)	(ow)	(prel)	(IN)	V	O	R	D	NC	NULL	PN	H
%	4%	5%	0%	0%	1%	0%	1%	1%	0%	0%	0%	1%	0%	1%	1%	3%	20%	23%	7%	21%	6%	3%	3%




# Some surprises

	Target	Error
→	Northwest	Earth less
	Northwesterner	Northwestern Scenario Northwest Eric
→	Me	Maine
→	Certain [sɪʔɪn]	*no error*
→	hooman	Whom
→	Jobs for	Javascript
	A lot of it	online

# Conclusions

- This research has accomplished a cross-ethnicity comparison of dialect-based ASR performance
  - Important! Quantified contribution of linguistic variables to error profile
- Is leveraging sociolinguistic knowledge of the fine phonetic detail in dialect variation worth it? *Yes!*
  - Eliminate approximately 26% of observed errors
- Worthwhile for linguists, too. ASR is a useful tool on the way to “actual” linguistic analysis.
- Not fast (sociophonetic analysis automated for vowels, not for consonants, not for non-majority dialects)
  
- Room for collaboration on transcription error reduction
- Room to improve access for people to services that rely increasingly upon ASR.

# Just for Fun...Top Ten Errors

	Error	Target
10.	pza	pa
9.	l zic	Isaac
8.	arndern	and during
7.	woon did	wounded
6.	Freycinet	A feast isn't it?
5.	anfang	fawn
4.	edgecator	educator
3.	plagge	plague
2.	Lenny Edge	lineage
1.	Grandpa Minecraft	Grandpa minded 

Thank you!

[wassink@uw.edu](mailto:wassink@uw.edu)

Slides: <https://depts.washington.edu/sociolab>

Try CLOx:

<https://clox.ling.washington.edu/>

## ***References***

diPaolo, M., and M. Yaeger-Dror (2011) *Sociophonetics: a student's guide*, London: Routledge.

Foulkes, P., Scobbie, J., and Watt, D. (2010) "Sociophonetics," in W. Hardcastle, J. Laver, and F. Gibbon (eds.), *Handbook of Phonetic Sciences*, 2<sup>nd</sup> ed., Oxford: Blackwell, 703-54.