

# Trade-offs in computational sociolinguistics methods

## Accuracy vs. fairness in forced alignment-based auto-coding

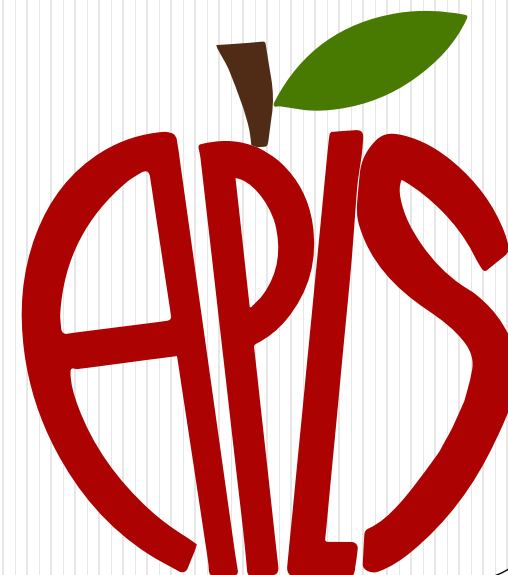
**Dan Villarreal** (*he/him*, [d.vill@pitt.edu](mailto:d.vill@pitt.edu))

University of Pittsburgh

Forced alignment-based auto-coding  
doesn't seem to work too good for coronal stop deletion 🙄

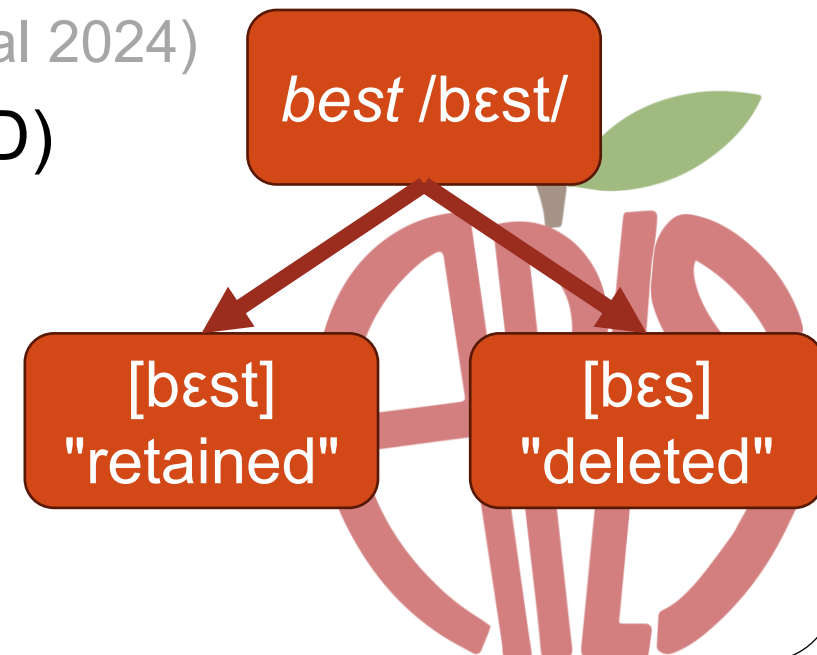
**Dan Villarreal** (*he/him*, [d.vill@pitt.edu](mailto:d.vill@pitt.edu))

University of Pittsburgh



# Sociolinguistic auto-coding (SLAC)

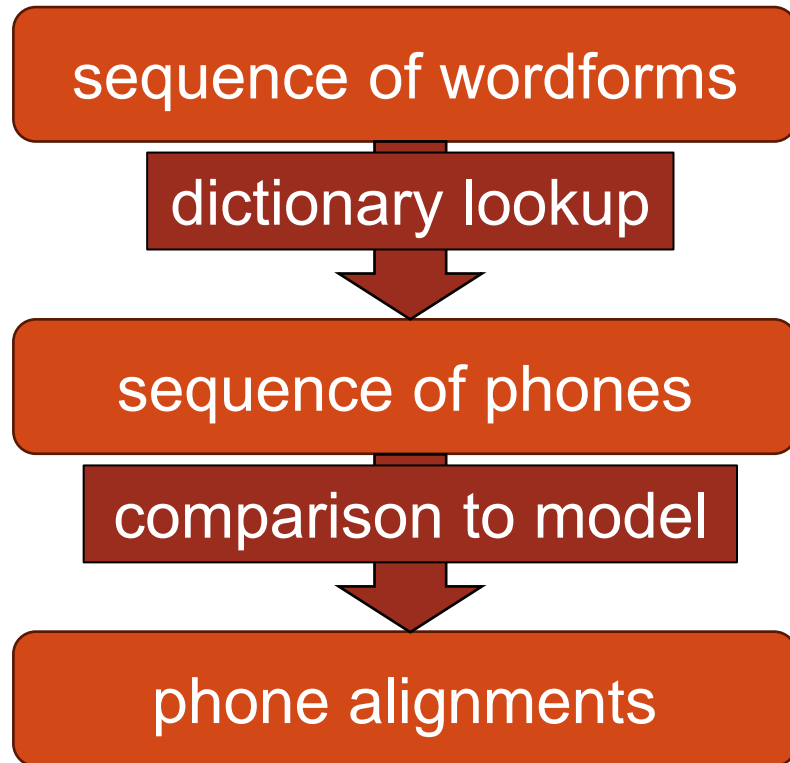
- SLAC: machine learning for labeling tokens with variants (Bailey 2016; McLarty et al. 2019; Villarreal et al. 2020; Kendall et al. 2021)
- But questions remain:
  - How to implement it
  - Whether it's meant to replicate human coding
  - How to account for intergroup fairness (Villarreal 2024)
- Today: SLAC with coronal stop deletion (CSD)



# SLAC via forced-alignment

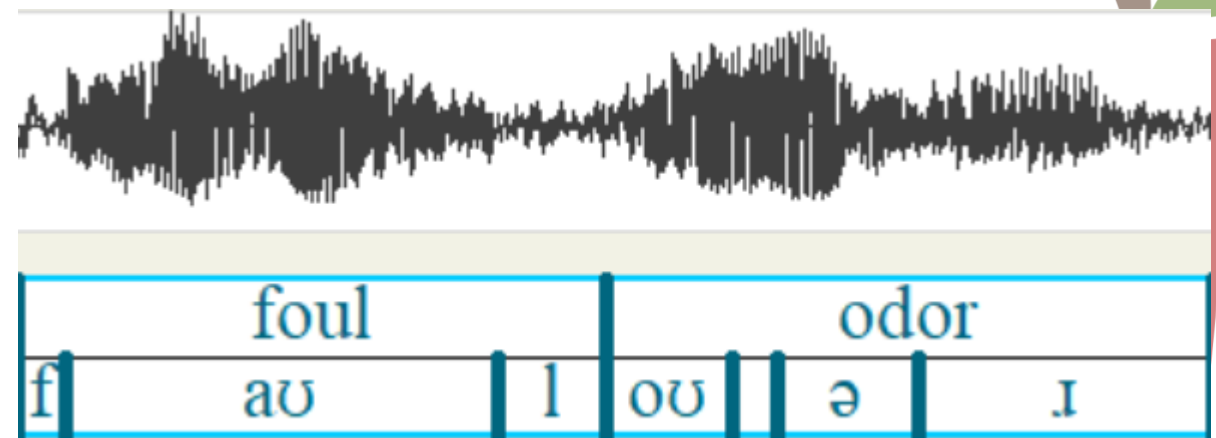
(Bailey 2016; Kendall et al. 2021)

- Exploits forced-aligners' ability to consider multiple candidate phonological representations for the same lexical item



foul odor

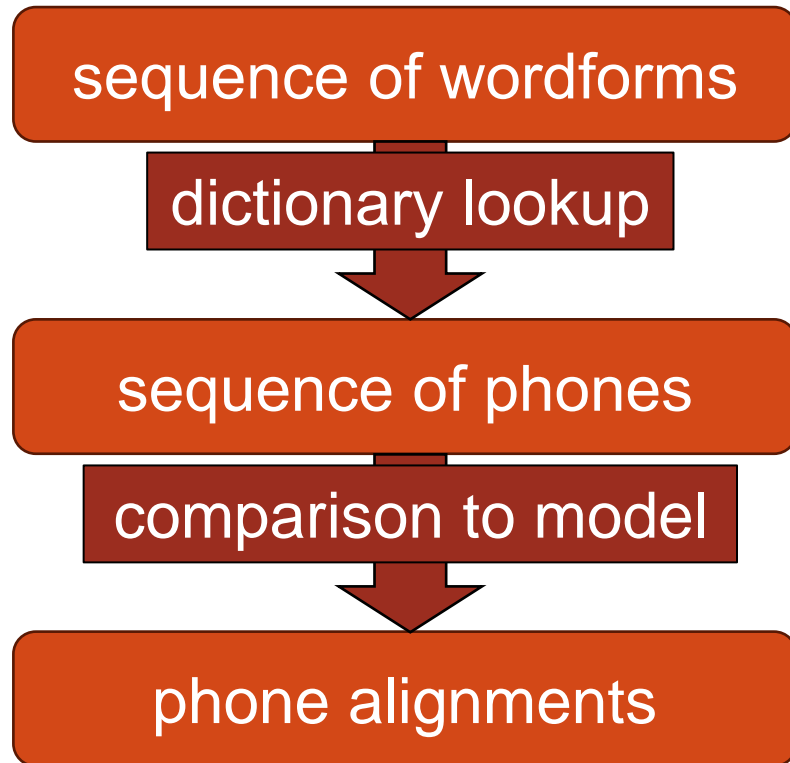
faʊl oʊdəʊ



# SLAC via forced-alignment

(Bailey 2016; Kendall et al. 2021)

- Exploits forced-aligners' ability to consider multiple candidate phonological representations for the same lexical item



**the odor**

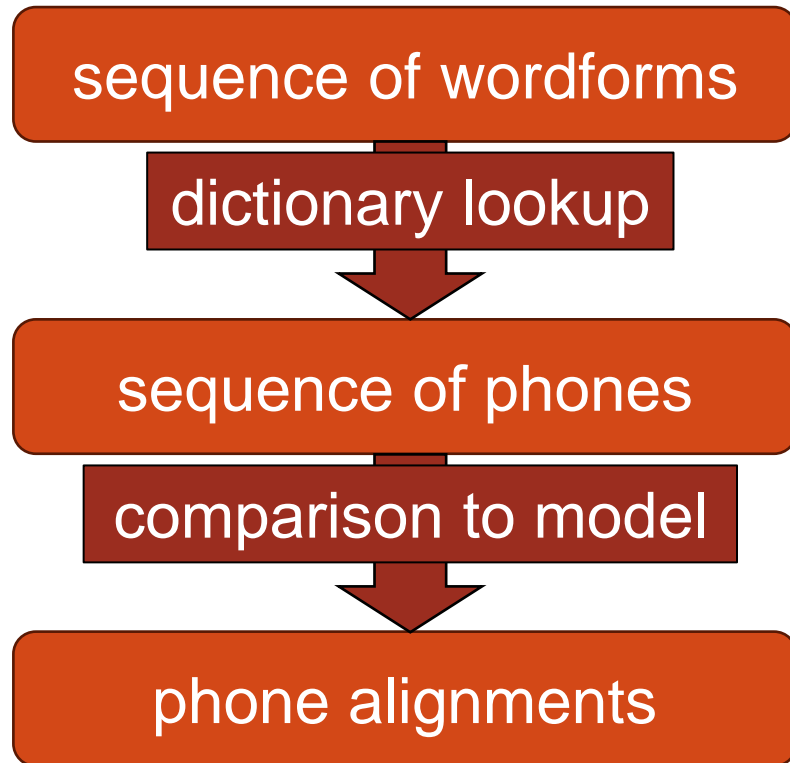
**(ðɪ | ðə)** ɒdəʊ



# SLAC via forced-alignment

(Bailey 2016; Kendall et al. 2021)

- Exploits forced-aligners' ability to consider multiple candidate phonological representations for the same lexical item



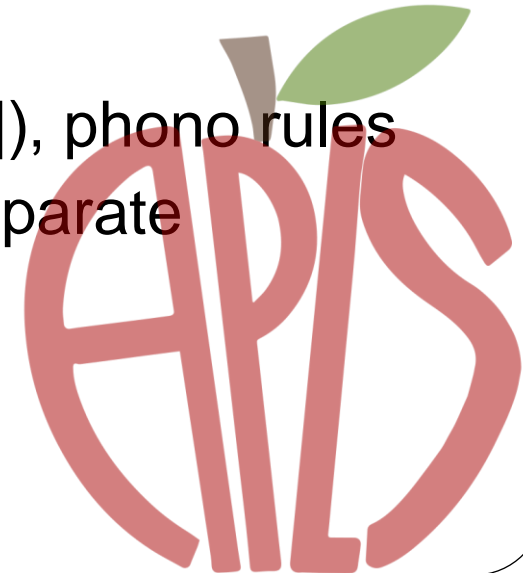
best odor

(best|bes) ʊdə



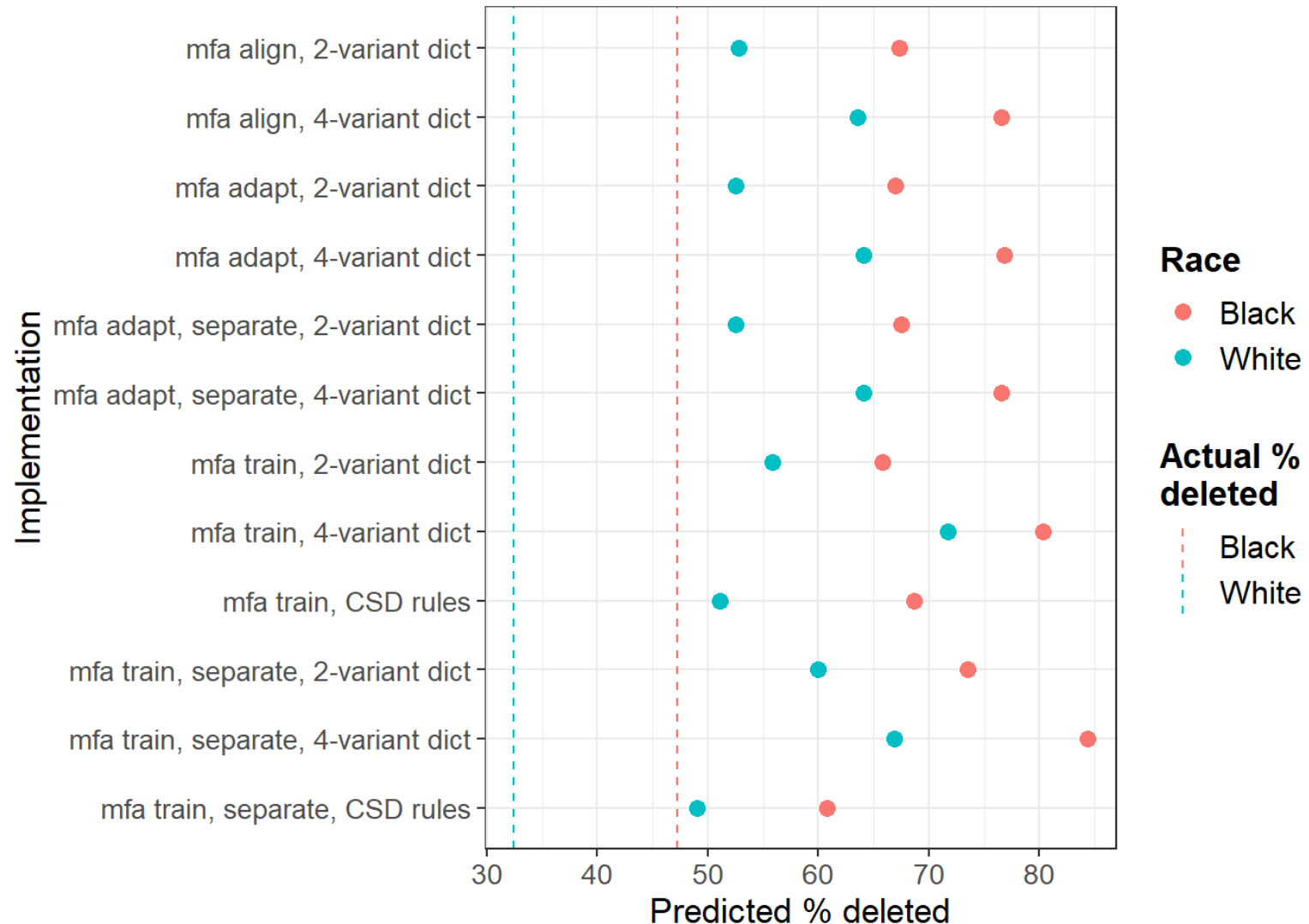
# Data & models

- 10,199 CSD tokens from Black & White speakers in Pittsburgh
  - 1005 hand-coded as retained (incl. [r ?]) or deleted
  - Black speakers: 47% deleted (240/508). White: 32% (161/497)
- Entire corpus (33 hours) aligned with Montreal Forced Aligner version 3.1.0 (McAuliffe et al. 2017), in 12 implementations (models):
  - Workflow: align with off-the-shelf `english_mfa` model, adapt `english_mfa`, train-and-align
  - Strategy: 2-variant dictionary, 4-variant dictionary (w/ [r ?]), phono rules
  - Dataset for train-and-align: corpus all at once vs. race-separate
- Output word alignments checked for final [t d (r ?)]



# Predictions for hand-coded data ( $n = 1005$ )

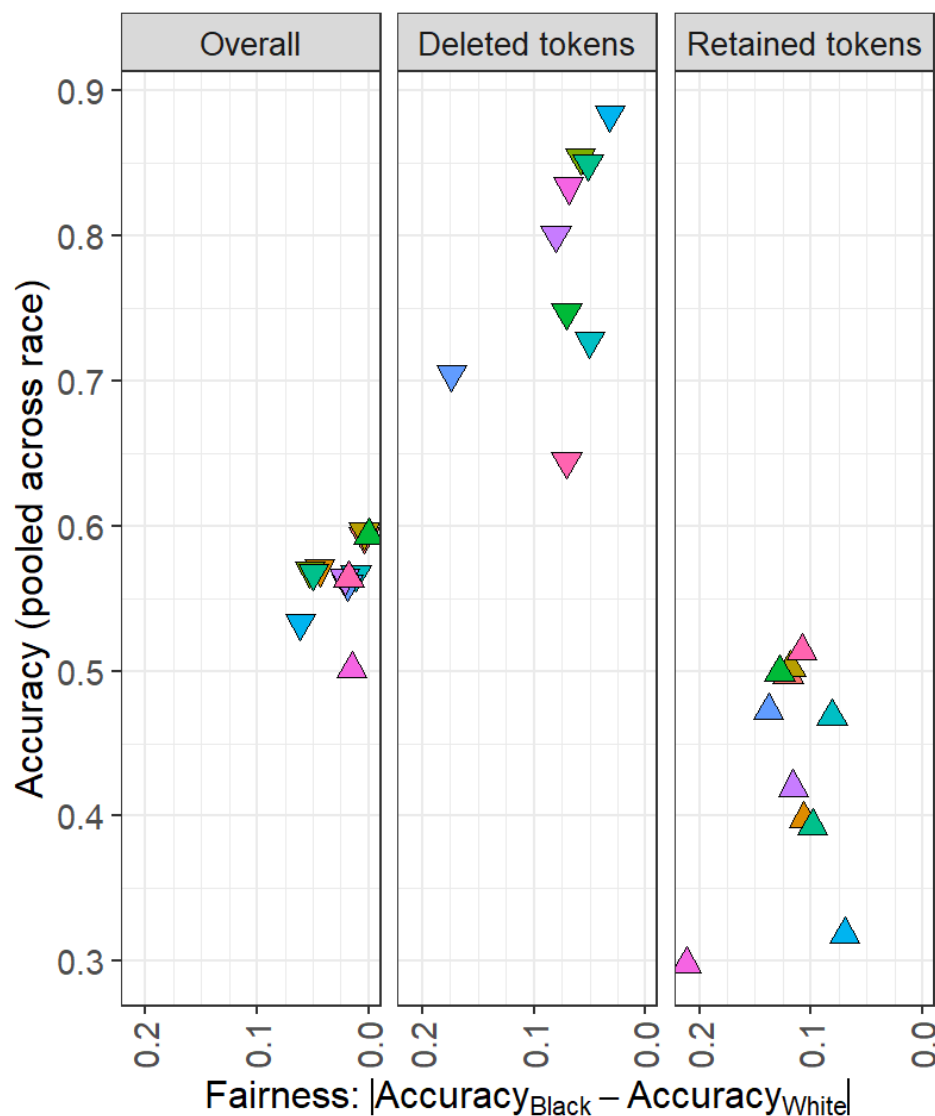
- All models **over-**predicted deletion
- 9/12 **under-**predicted magnitude of race effect on deletion
  - Range of predicted diffs: 8.7–17.6pp
- Difficulty of detecting [f̥ tʔ r ʔ]?





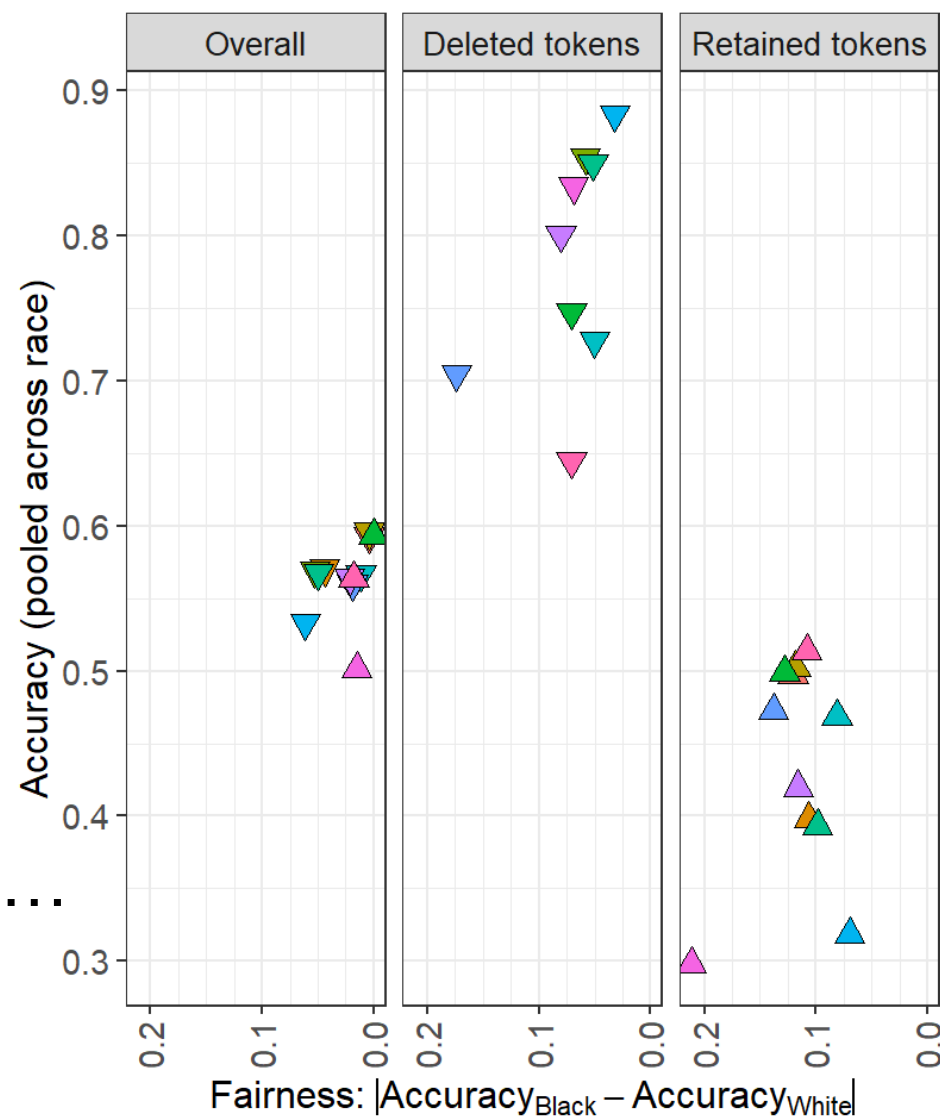
# Accuracy

- SOTA: ~85%
- Here: 50.1–59.6%
  - Worse for train-and-align
- Better accuracy for (actual) deleted than retained
  - Because deleted over-predicted



# Fairness

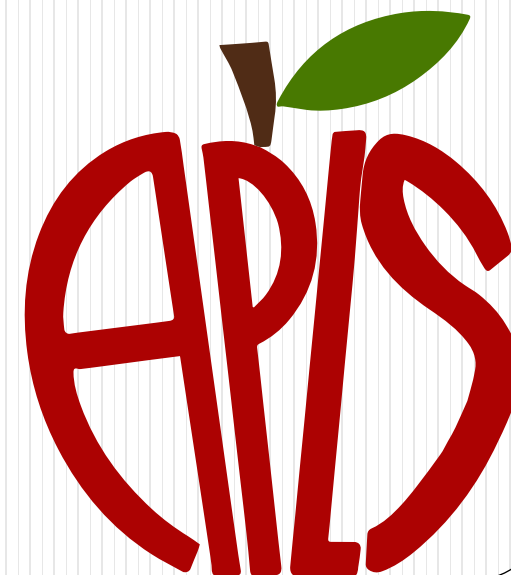
- $|\text{Acc}_{\text{Black}} - \text{Acc}_{\text{White}}|$
- Overall: 0.02–6pp
- Among deleted tokens: 3–17pp
  - Worse for White
- Among retained tokens: 7–21pp
  - Worse for Black
- Hence my takeaway...



Forced alignment-based auto-coding  
doesn't seem to work too good for coronal stop deletion 🙄

**Dan Villarreal** (*he/him*, [d.vill@pitt.edu](mailto:d.vill@pitt.edu))

University of Pittsburgh



# Introducing APLS

- The Archive of Pittsburgh Language and Speech (APLS): a soon-to-be publicly accessible corpus of Pittsburgh socioLx interviews

Journal of English Linguistics



Restricted access | Research article | First published June 2006

Mobility, Indexicality, and the Enregisterment of “Pittsburghese”

[Barbara Johnstone](#), [Jennifer Andrus](#), and [Andrew E. Danielson](#) [View all authors and affiliations](#)

[Volume 34, Issue 2](#) | <https://doi.org/10.1177/0075424206290692>

*Barbara Johnstone, Daniel Baumgardt,  
Maeve Eberhardt, Scott Kiesling*  
**PITTSBURGH SPEECH  
AND PITTSBURGHESSE**

ES IN SOCIOLINGUISTICS

BARBARA JOHNSTONE

**SPEAKING  
PITTSBURGHESSE**  
THE STORY OF A DIALECT

CHAPTER 12

AFRICAN AMERICAN  
LANGUAGE IN  
PITTSBURGH  
AND THE LOWER  
SUSQUEHANNA VALLEY

JENNIFER BLOOMQUIST AND SHELOME GOODEN

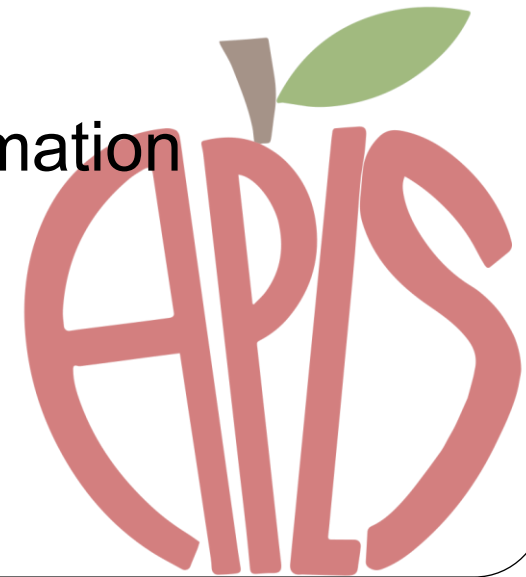
Language & Communication  
[Volume 32, Issue 4](#), October 2012, Pages 358-371

Enregisterment of Pittsburghese and  
the local African American community

Maeve Eberhardt

# Introducing APLS

- The Archive of Pittsburgh Language and Speech (APLS): a soon-to-be publicly accessible corpus of Pittsburgh socioLx interviews
- Powered by LaBB-CAT corpus management framework (Fromont & Hay 2012)
- Currently: 33 hours of speech from 34 interviewees
  - 387K word tokens, 956K aligned segments
  - When complete: 45 hours from 40 interviewees
- Richly annotated with multiple layers of linguistic information
- Example: searching CSD tokens for speaker "HD06"



# Thanks! Questions?

Thanks to Scott Kiesling, Barbara Johnstone, Robert Fromont, members of the Computational Sociolinguistics Lab past and present, and—most of all—the speakers who have shared their voices with us

**Dan Villarreal** (*he/him*, [d.vill@pitt.edu](mailto:d.vill@pitt.edu))  
University of Pittsburgh

