Investigating morphosyntactic variation in African American English on Twitter

Tessa Masis[1]
they/them/theirs

Taylor Jones[2]
he/him/his

Chloe Eggleston[1]
she/her/hers

Meghan Armstrong[1]
she/her/hers

Lisa Green[1]
she/her/hers

Brendan O'Connor[1]
he/him/his

[1]University of Massachusetts Amherst

[2]Naval Postgraduate School

Overview

- Research questions
  - Uniformity/variation within AAE
- Data & approach
  - Corpus of 227M tweets
  - Automatically detecting morphosyntactic features
- Results
  - Regional variation
  - Demographic variation
- Conclusion

Variation in AAE

Sociolinguistic Folklore in the Study of African American English

Walt Wolfram*
North Carolina State University

REGIONALITY IN THE DEVELOPMENT OF AFRICAN AMERICAN ENGLISH

WALT WOLFRAM AND MARY E. KOHN

A focus on a core set of basilectal structures in non-Southern urban communities obscured regional variation in early sociolinguistic studies of African American English (AAE). However, community comparisons, particularly in the rural South, indicate that regionality has played an essential role in the past and present development of the variety. This current analysis compares apparent time evidence for

Variation in AAE

Sociolinguistic Folklore in the Study of African American English

Walt Wolfram*
North Carolina State University

REGIONALITY IN THE DEVELOPMENT OF AFRICAN AMERICAN ENGLISH

WALT WOLFRAM AND MARY E. KOHN

Yaeger-Dror (2007), Wroblewski et al. (2009), Yaeger-Dror & Thomas (2010), Lee (2016), Austen (2017), Jones (2020)

Research questions

- To what extent is there **systematic morphosyntactic variation within AAE**?

- How much of this variation can be **accounted for by social factors** (i.e. region, race, age, socioeconomic status)?

Data

- 227M geotagged tweets from Twitter Gardenhose
- Posted from the US during May 2011 - April 2015
- Filtered to prioritize conversational language and limit automated posts

- 5 orders of magnitude larger than previous Twitter corpus studies of AAE, with at least some data in all US counties

Morphosyntactic features

Feature	Example
*Zero possessive	*they want to do they own thing*
Overt possessive	*they want to do their own thing*
*Zero copula	*she the folk around here*
Overt copula	*she's the folk around here*
*future *gone*	*we gone rock it out like*
*Habitual *be*	*I just be liking the beat*
*Resultant *done*	*you done lost your mind*
**be done*	*I be done died walking up that many*
**steady*	*and you steady talking to them*
**finna*	*she's finna have a baby*
*Negative concord	*I ain't doing nothing wrong*
Single negative	*I ain't doing anything wrong*
*Negative auxiliary inversion	*nobody don't say nothing*
*Preverbal negator *ain't*	*I ain't doing nothing wrong*
*Zero 3rd person singular present tense *-s*	*I don't know if it count*
**is/was* generalization	*they is die hard Laker fans*
*Double-object construction	*I got me my own car*
**Wh*-question	*what they were doing?*

Automatic feature detection

- Task: given a set of features F, for each f $\in$ F identify utterances which contain f

- For our large dataset, automatic methods are a valuable alternative to manual annotation

Automatic feature detection: our framework

- Generate a small contrast set

- Fine-tune BERT on this contrast set, where each head is a binary classifier for a single feature

Automatic feature detection: our framework

- Generate a small contrast set
    - A labeled collection of positive and negative examples that are highly similar, where a positive example has the feature/label and a negative example does not (Gardner et al. 2020)

I be out at my bus stop every day.

I’m out at my bus stop every day.
I’ll be out at my bus stop every day.
I would be out at my bus stop every day.

Automatic feature detection: our framework

- Generate a small contrast set

Corpus-Guided Contrast Sets for Morphosyntactic Feature Detection in Low-Resource English Varieties

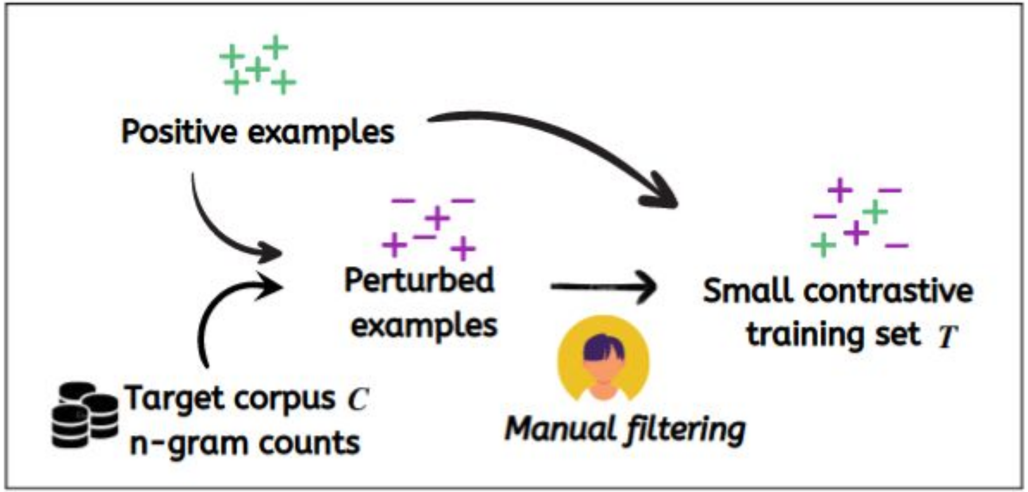
Tessa Masis *they/them/theirs*
Anissa Neal *she/her/hers*
Lisa Green *she/her/hers*
Brendan O'Connor *he/him/his*

University of Massachusetts Amherst
{tmasis,brenocon}@cs.umass.edu
{anneal,lgreen}@linguist.umass.edu

Field Matters @ COLING2022

CGEdit

- Input:
  - Seed set of positive examples
  - Target corpus n-gram counts



- Method:
  - Corpus-guided edits
  - Human-in-the-loop filtering
- Output:
  - Morphosyntactically contrastive training data

Automatic feature detection: our framework

- Generate a small contrast set

- Fine-tune BERT on this contrast set, where each head is a binary classifier for a single feature

Automatically detecting features

- Input: 227M geotagged tweets

- Output: Census tract-level relative frequencies for 18 morphosyntactic features

$$rf_{feat} = \#\ \text{tweets with feature} \,/\, \#\ \text{total tweets}$$

Automatically detecting features

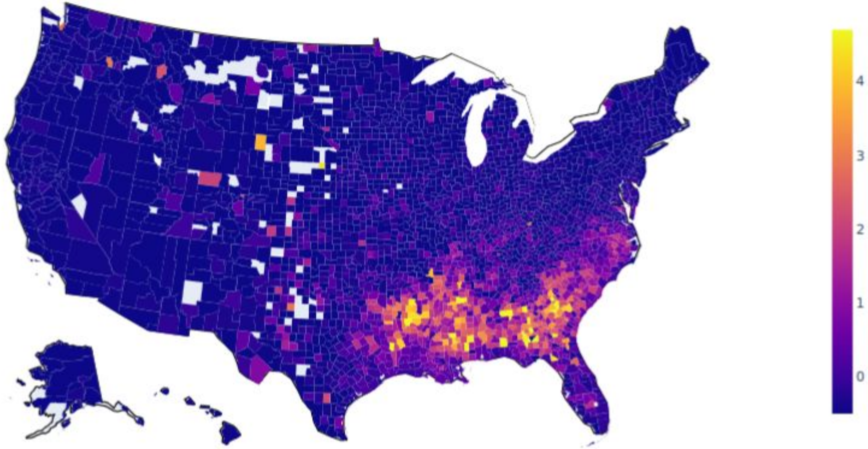
- Input: 227M geotagged tweets

- Output: Census tract-level relative frequencies for 18 morphosyntactic features

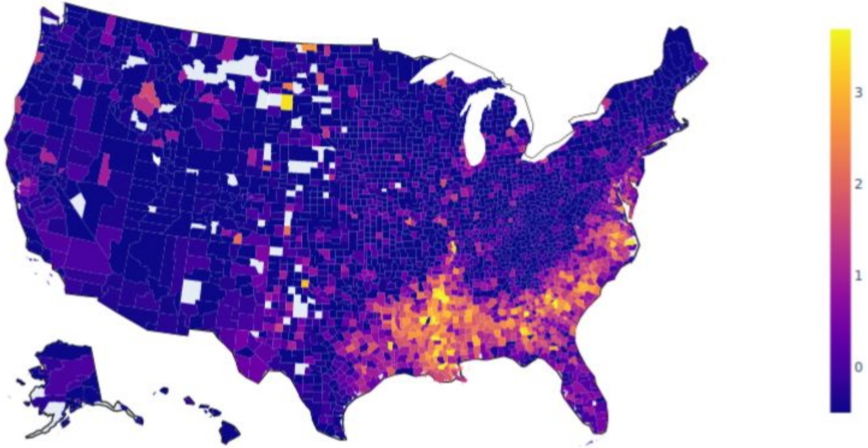
$$rf_{feat} = \text{\# tweets with feature} \, / \, \text{\# total tweets}$$

$$z_{feat} = (rf_{feat} - \mu_{feat}) \, / \, \sigma_{feat}$$

Automatically detecting features

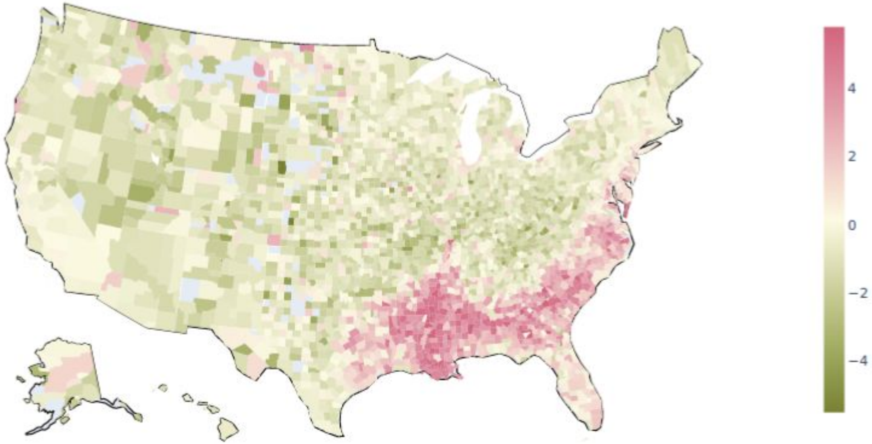


(a) Distribution of resultant *done*

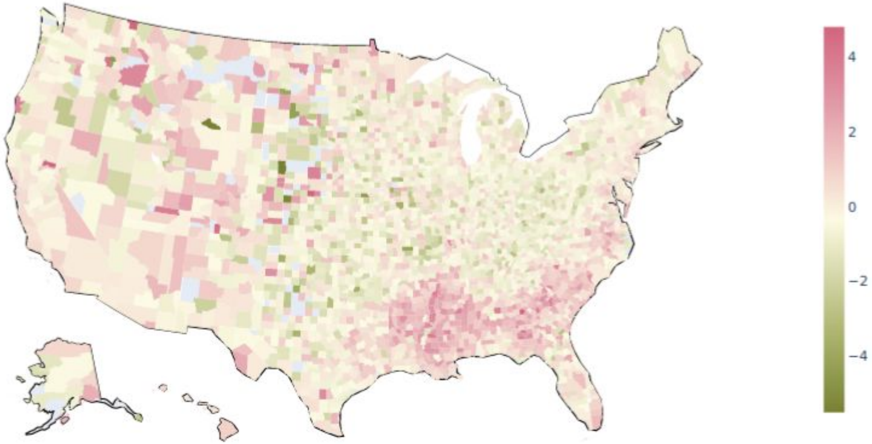


(b) Distribution of habitual *be*

Automatically detecting features



(c) Distribution of zero copula versus overt copula



(d) Distribution of negative concord versus single negative

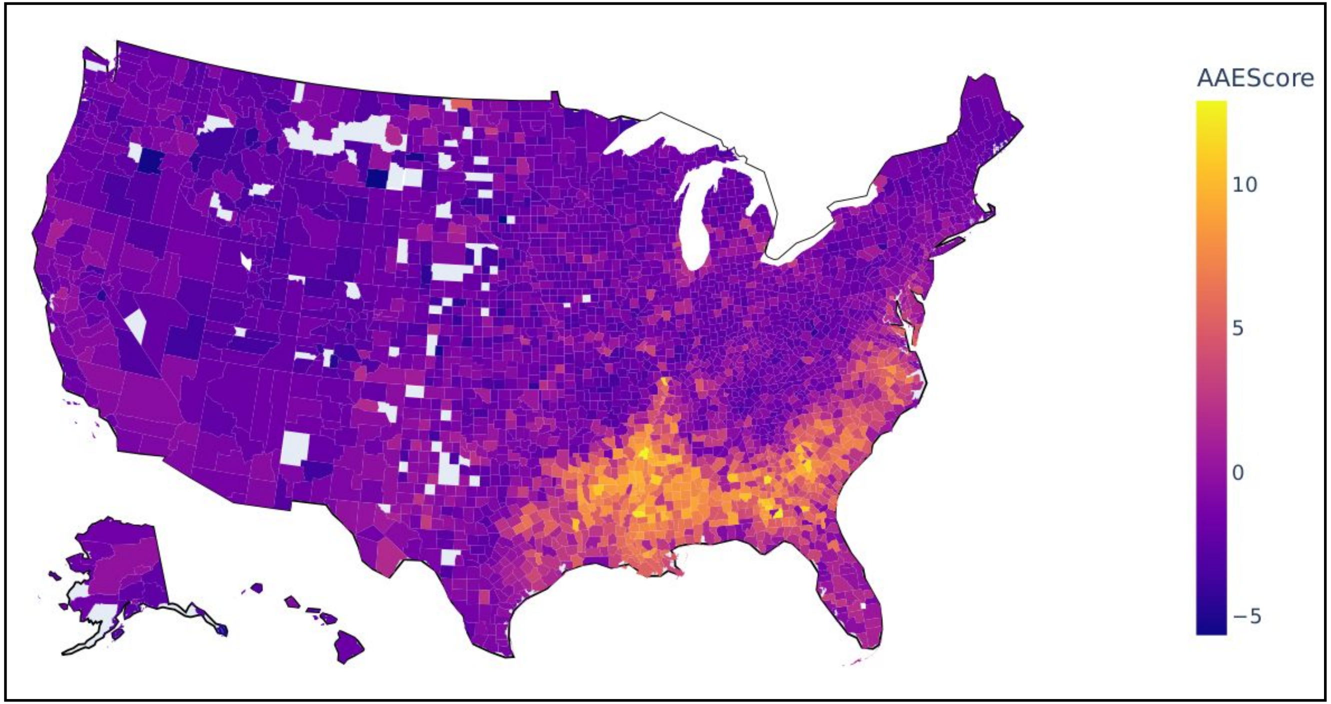
Research questions

- To what extent is there **systematic morphosyntactic variation within AAE**?
    - Principal Components Analysis (PCA)

PCA: feature loadings

Feature	Frequency	AAEScore
ain't	2,168,105	.9156
Habitual *be*	947,900	.8436
future *gone*	477,514	.8409
Negative concord	1,473,423	.8258
Zero copula	7,726,637	.7867
Zero 3rd person singular present tense *-s*	1,100,333	.6721
finna	769,822	.6261
Negative auxiliary inversion	135,497	.6106
Resultant *done*	86,933	.5794
Wh-question	1,517,957	.5754
Zero possessive	239,302	.4587
Double object	486,346	.3767
Single negative	22,907,646	.3037
is/was generalization	1,321,730	.2814
steady	15,047	.2248
be done	146	.0509
Overt possessive	2,735,250	-.4840
Overt copula	53,925,152	-.7126
Percentage of variance		35.58

PCA: AAEScore



Research questions

- To what extent is there **systematic morphosyntactic variation within AAE**?
    - Principal Components Analysis (PCA)

- How much of this variation can be **accounted for by social factors** (i.e. region, race, age, socioeconomic status)?
    - Correlation analysis
    - Linear regression

Correlation analysis

	Pearson's r
Afr.-Am. pop.	0.79
RUCA	-0.07
Latitude	-0.24
Mexican pop.	-0.04
PR pop.	0.07
Income	-0.39
…	…

Linear Regression analysis: RUCA

	Pearson's r	(1)
Afr.-Am. pop.	0.79	2.07
RUCA	-0.07	0.06
Latitude	-0.24	
Mexican pop.	-0.04	
PR pop.	0.07	
Income	-0.39	
…	…	

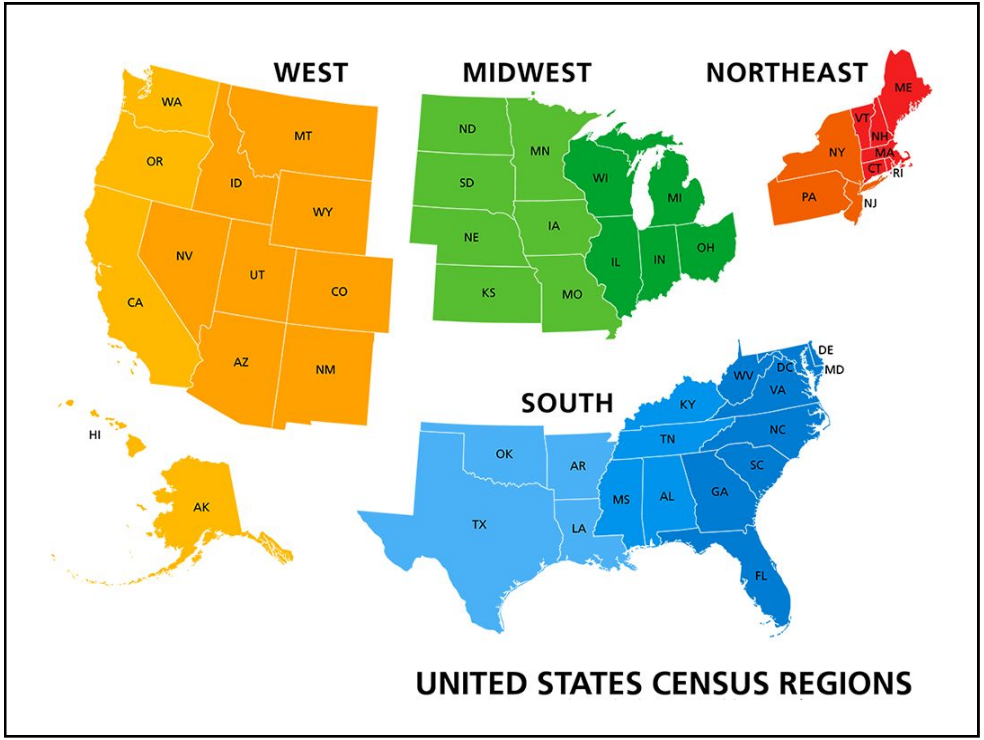
Linear Regression analysis: RUCA + latitude

	Pearson's r	(1)	(2)
Afr.-Am. pop.	0.79	2.07	2.03
RUCA	-0.07	0.06	0.09
Latitude	-0.24		-0.40
Mexican pop.	-0.04		
PR pop.	0.07		
Income	-0.39		
…	…		

Linear Regression analysis: Mexican pop.

	Pearson's r	(1)	(2)	(3)
Afr.-Am. pop.	0.79	2.07	2.03	2.09
RUCA	-0.07	0.06	0.09	
Latitude	-0.24		-0.40	
Mexican pop.	-0.04			0.19
PR pop.	0.07			
Income	-0.39			
...	...			

Rural South



WEST
MIDWEST
NORTHEAST
SOUTH
WA
MT
OR
ID
WY
NV
UT
CO
CA
AZ
NM
HI
AK
ND
MN
SD
WI
MI
NE
IA
IL
IN
OH
KS
MO
ME
VT
NH
MA
NY
CT
RI
PA
NJ
DE
MD
DC
WV
VA
KY
NC
TN
OK
AR
SC
GA
MS
AL
TX
LA
FL
UNITED STATES CENSUS REGIONS

Conclusions

- To what extent is there **systematic morphosyntactic variation within AAE**?
  - There is systematic variation, which can be characterized by our first principal component (AAEScore)

- How much of this variation can be **accounted for by social factors** (i.e. region, race, age, socioeconomic status)?
  - Can mostly be explained by relative African American population; but urbanization, geographic region, racial identity also play a role

Thank you!

Slides and abstract available at
tmasis.github.io/

This material is based upon work supported by the National Science Foundation under grant BCS-2042939. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Tessa Masis
tmasis@cs.umass.edu

Chloe Eggleston
ceggleston@umass.edu

Lisa Green
lgreen@linguist.umass.edu

Taylor Jones
thelanguagejones@gmail.com

Meghan Armstrong
armstrong@spanport.umass.edu

Brendan O'Connor
brenocon@cs.umass.edu