A large-scale Twitter-based exploration of morphosyntactic geographic variation in African American English

Tessa Masis[1]
they/them/theirs

Chloe Eggleston[1]
she/her/hers

Anissa Neal[2]
she/her/hers

Lisa Green[2]
she/her/hers

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

[1]College of Information and Computer Sciences
University of Massachusetts Amherst

[2]Department of Linguistics
University of Massachusetts Amherst

Overview

- Research questions
  - Uniformity/variation within AAE

Overview

- Research questions
  - Uniformity/variation within AAE
- Data & approach
  - Corpus of 224M tweets
  - Automatically detecting morphosyntactic features

Overview

- Research questions
  - Uniformity/variation within AAE
- Data & approach
  - Corpus of 224M tweets
  - Automatically detecting morphosyntactic features
- Preliminary analysis
  - Regional variation
  - Feature co-occurrence

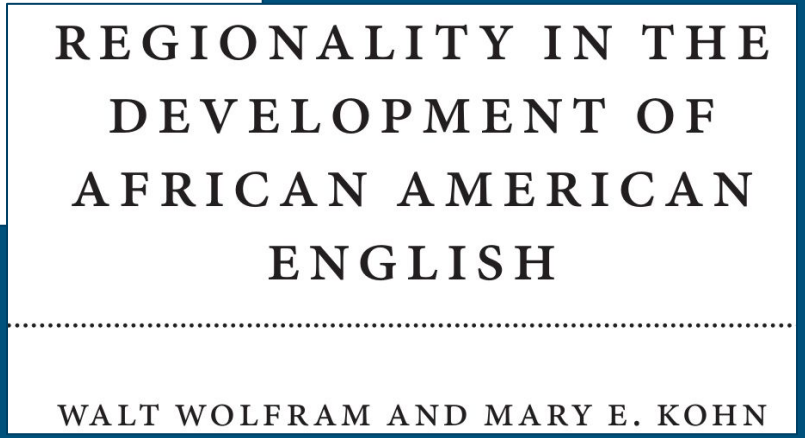
Overview

- Research questions
  - Uniformity/variation within AAE
- Data & approach
  - Corpus of 224M tweets
  - Automatically detecting morphosyntactic features
- Preliminary analysis
  - Regional variation
  - Feature co-occurrence
- Future directions

Uniformity in AAE?

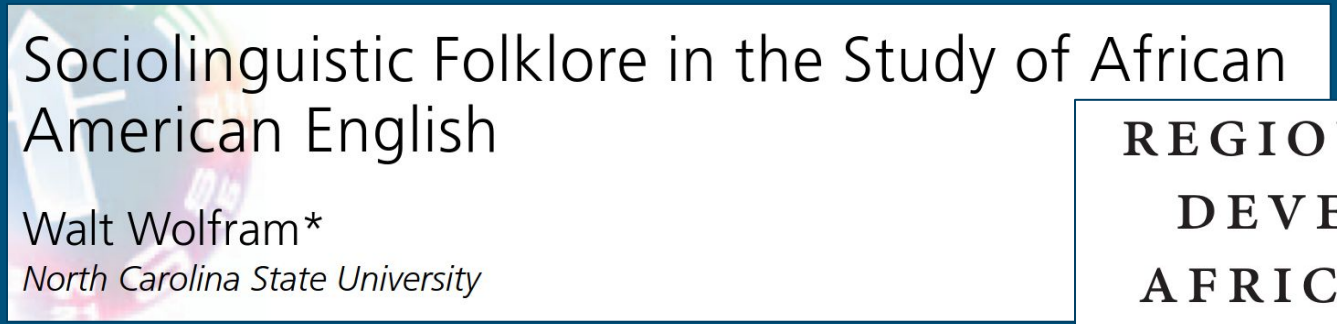
Sociolinguistic Folklore in the Study of African American English

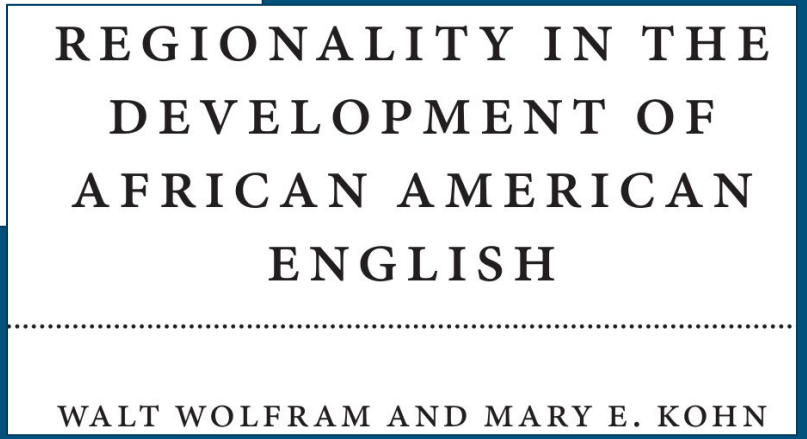
Walt Wolfram*
North Carolina State University



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





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

Research questions

- To what extent is there **uniformity** and/or **systematic variation within AAE**?

Research questions

- To what extent is there **uniformity** and/or **systematic variation within AAE**?

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

Data

- 224M geotagged tweets from Twitter Decahose
- 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 utterance
Zero possessive -*'s*	go over my grandmama house
Overt possessive -*'s*	go over my grandmama's house
Zero copula	she _ the folk around here
Overt copula	she is the folk around here
gone	we gone rock it out like
Habitual *be*	I just be liking the beat
Remote past stressed *BIN*	but I BEEN having that one
Resultant *done*	you done lost your mind
Habitual *be* + resultant *done*	so they be done gone to school
Stressed *BIN* + resultant *done*	he BEEN done put that in there
steady	and you steady talking to them
finna	she's finna have a baby
Double modal	he might could really get our minds
Negative concord	I ain't doing nothing wrong
Single negative	I ain't doing anything wrong
Negative auxiliary inversion	don't nobody know what I had
Non-inverted negative concord	nobody don't say nothing
Preverbal negator *ain't*	I ain't doing nothing wrong
Zero 3rd p sg present tense -*s*	I don't know if it count
Narrative/habitual -*s*	so I gets in the car
is/was-generalization	they is die hard Laker fans
Zero plural -*s*	about four or five month
Double-object construction	I got me my own car
Wh-question	what they was doing?

Many of the AAE-specific features selected from Green (2002) and Koenecke et al. (2020)

Morphosyntactic features

Feature	Example utterance
Zero possessive *-'s*	go over my grandmama house
Overt possessive *-'s*	go over my grandmama's house
Zero copula	she _ the folk around here
Overt copula	she is the folk around here
gone	we gone rock it out like
Habitual *be*	I just be liking the beat
Remote past stressed *BIN*	but I BEEN having that one
Resultant *done*	you done lost your mind
Habitual *be* + resultant *done*	so they be done gone to school
Stressed *BIN* + resultant *done*	he BEEN done put that in there
steady	and you steady talking to them
finna	she's finna have a baby
Double modal	he might could really get our minds
Negative concord	I ain't doing nothing wrong
Single negative	I ain't doing anything wrong
Negative auxiliary inversion	don't nobody know what I had
Non-inverted negative concord	nobody don't say nothing
Preverbal negator *ain't*	I ain't doing nothing wrong
Zero 3rd p sg present tense *-s*	I don't know if it count
Narrative/habitual *-s*	so I gets in the car
is/was-generalization	they is die hard Laker fans
Zero plural *-s*	about four or five month
Double-object construction	I got me my own car
Wh-question	what they was doing?

Morphosyntactic features

Feature	Example utterance
Zero possessive *-'s*	go over my grandmama house
Overt possessive *-'s*	go over my grandmama's house
Zero copula	she _ the folk around here
Overt copula	she is the folk around here
gone	we gone rock it out like
Habitual *be*	I just be liking the beat
Remote past stressed *BIN*	but I BEEN having that one
Resultant *done*	you done lost your mind
Habitual *be* + resultant *done*	so they be done gone to school
Stressed *BIN* + resultant *done*	he BEEN done put that in there
steady	and you steady talking to them
finna	she's finna have a baby
Double modal	he might could really get our minds
Negative concord	I ain't doing nothing wrong
Single negative	I ain't doing anything wrong
Negative auxiliary inversion	don't nobody know what I had
Non-inverted negative concord	nobody don't say nothing
Preverbal negator *ain't*	I ain't doing nothing wrong
Zero 3rd p sg present tense *-s*	I don't know if it count
Narrative/habitual *-s*	so I gets in the car
is/was-generalization	they is die hard Laker fans
Zero plural *-s*	about four or five month
Double-object construction	I got me my own car
Wh-question	what they was doing?

'Principle of accountability' (Labov 1972; Tagliamonte 2006)

Morphosyntactic features

Feature	Example utterance
Zero possessive *-'s*	go over my grandmama house
Overt possessive *-'s*	go over my grandmama's house
Zero copula	she _ the folk around here
Overt copula	she is the folk around here
gone	we gone rock it out like
Habitual *be*	I just be liking the beat
Remote past stressed *BIN*	but I BEEN having that one
Resultant *done*	you done lost your mind
Habitual *be* + resultant *done*	so they be done gone to school
Stressed *BIN* + resultant *done*	he BEEN done put that in there
steady	and you steady talking to them
finna	she's finna have a baby
Double modal	he might could really get our minds
Negative concord	I ain't doing nothing wrong
Single negative	I ain't doing anything wrong
Negative auxiliary inversion	don't nobody know what I had
Non-inverted negative concord	nobody don't say nothing
Preverbal negator *ain't*	I ain't doing nothing wrong
Zero 3rd p sg present tense *-s*	I don't know if it count
Narrative/habitual *-s*	so I gets in the car
is/was-generalization	they is die hard Laker fans
Zero plural *-s*	about four or five month
Double-object construction	I got me my own car
Wh-question	what they was doing?

'Group orientation' (Alim & Reyes 2011)

Approach: automatically detecting features

- Task: given textual data, detect specific morphosyntactic features

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

Approach: automatically detecting features

- Generate a small contrast set

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

Approach: automatically detecting features

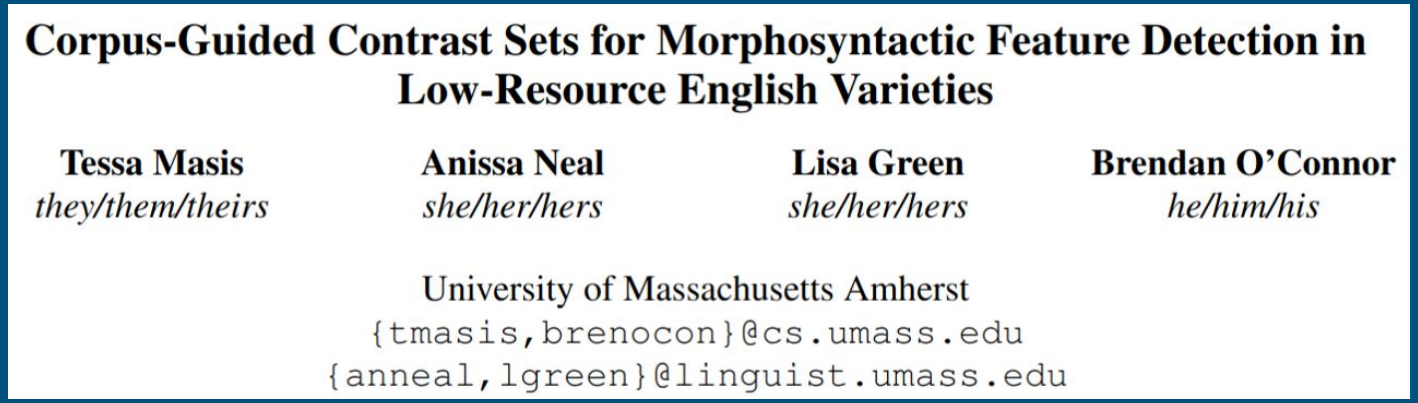
- Generate a small contrast set
  - A labeled collection of positive and negative examples that are highly similar, where a positive example has the feature/label and a negative example does not (Gardner et al., 2020)

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.

Approach: automatically detecting features

- 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

Approach: automatically detecting features

- Generate a small contrast set

- Fine-tune BERT on this contrast set, where each head is a binary classifier for a single feature
  - BERT: a large pretrained language model (Devlin et al., 2019)
  - Fine-tuning: taking a model trained on a large unlabeled dataset and doing partial retraining of it on a smaller labeled dataset for a downstream task

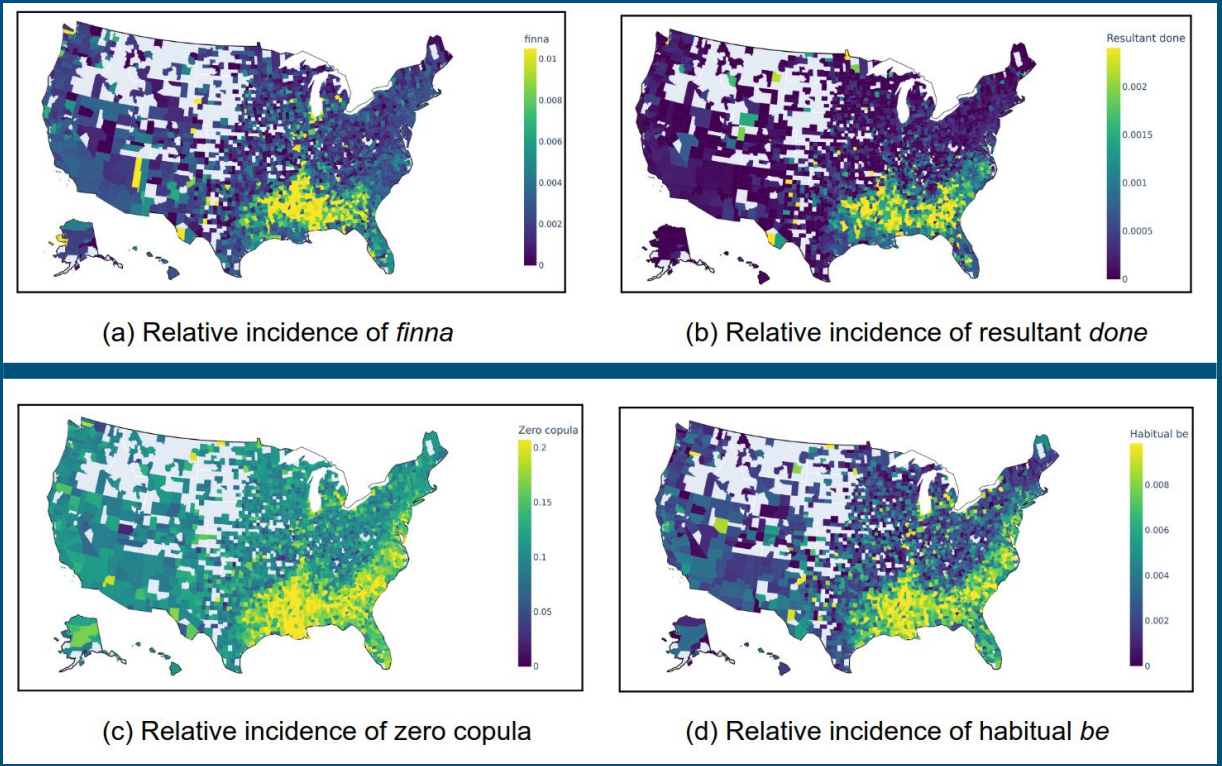
Approach: automatically detecting features

- Input: 224M geotagged tweets

- Output: County-level relative incidences for 24 morphosyntactic features

```
Relative incidence (feature) =
# tweets with feature / # total tweets
```

Preliminary analysis: regional variation



(a) Relative incidence of *finna*

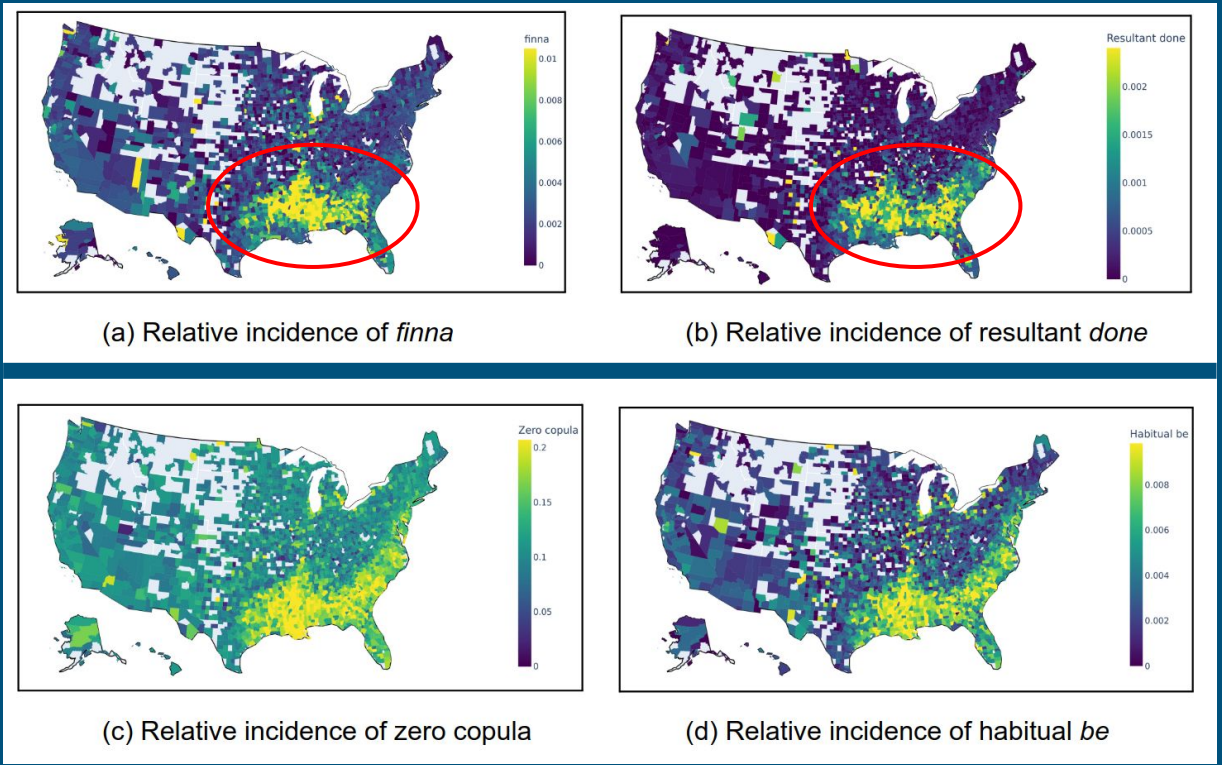
(b) Relative incidence of resultant *done*

(c) Relative incidence of zero copula

(d) Relative incidence of habitual *be*

Two morphosyntactic dialect regions

Preliminary analysis: regional variation



(a) Relative incidence of *finna*

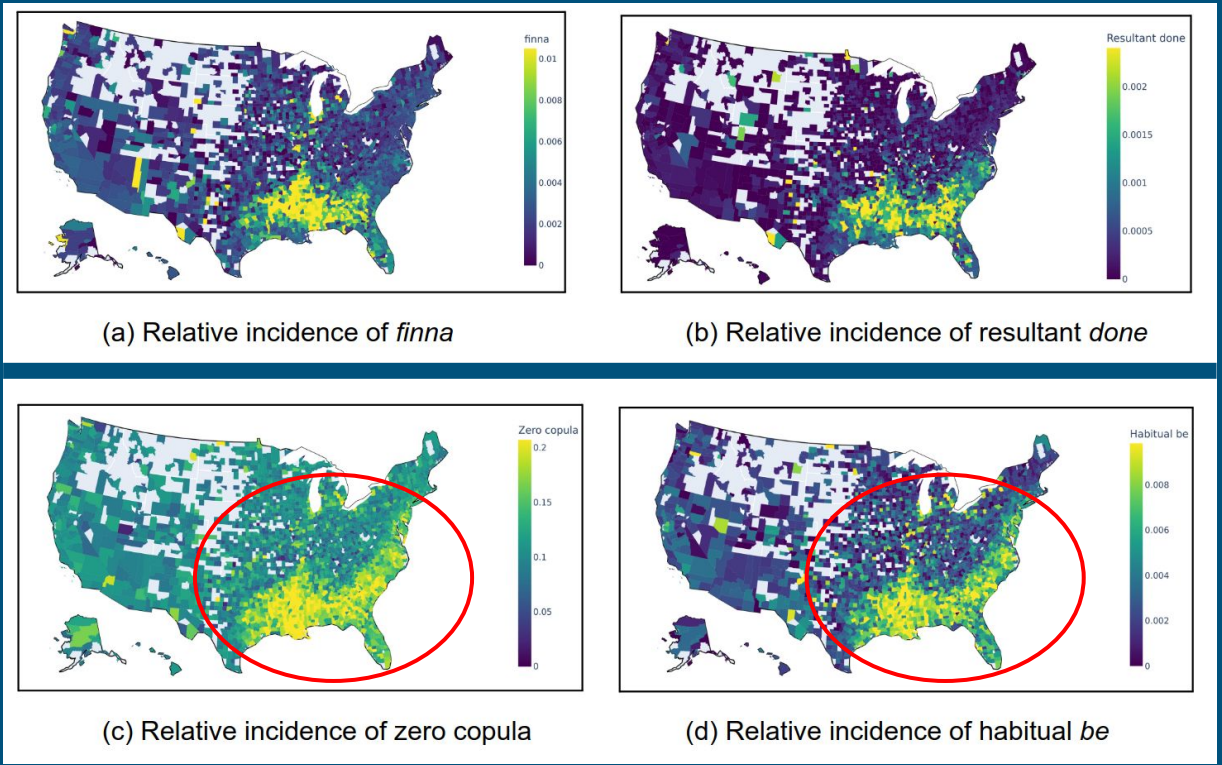
(b) Relative incidence of resultant *done*

(c) Relative incidence of zero copula

(d) Relative incidence of habitual *be*

Two morphosyntactic dialect regions

Preliminary analysis: regional variation



(a) Relative incidence of *finna*

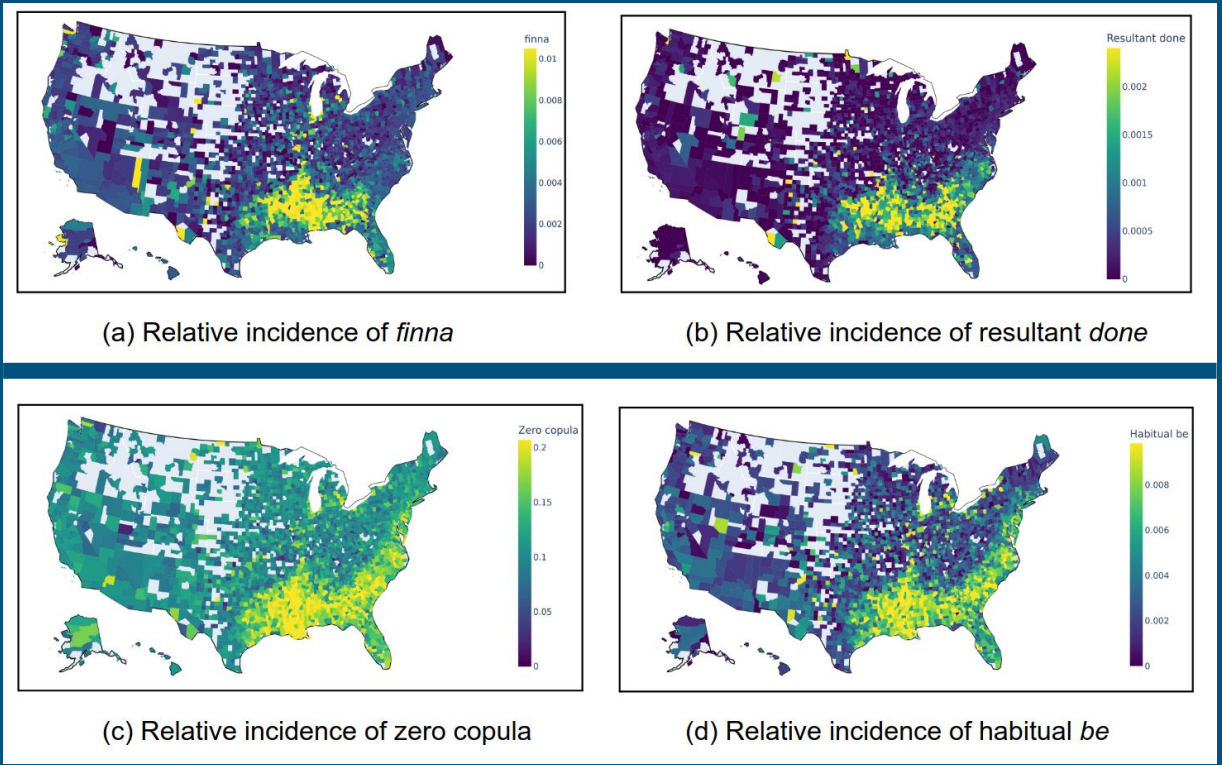
(b) Relative incidence of resultant *done*

(c) Relative incidence of zero copula

(d) Relative incidence of habitual *be*

Two morphosyntactic dialect regions

Preliminary analysis: regional variation



(a) Relative incidence of *finna*

(b) Relative incidence of resultant *done*

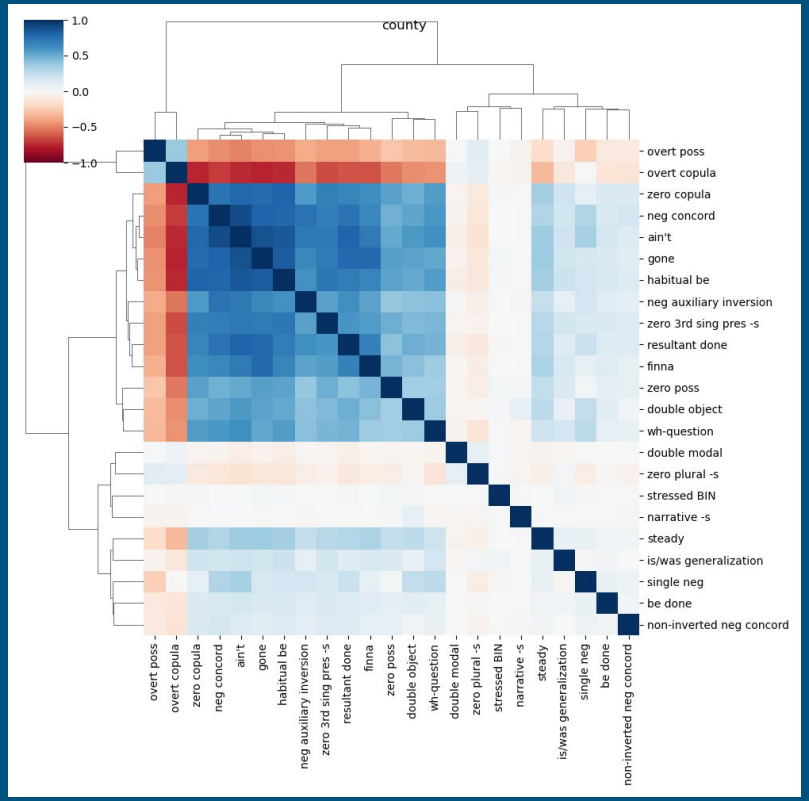
(c) Relative incidence of zero copula

(d) Relative incidence of habitual *be*

Two morphosyntactic dialect regions

Aligns with phonological and lexical variation in AAE (Jones 2015; Austen 2017; Jones 2020)

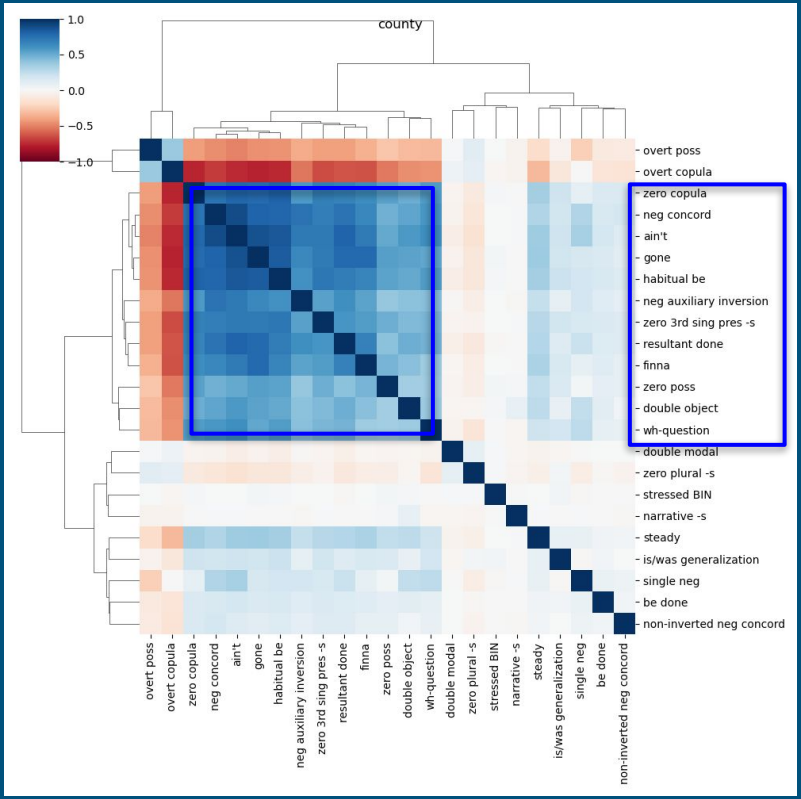
Preliminary analysis: feature co-occurrence



Feature-to-feature correlation heatmap

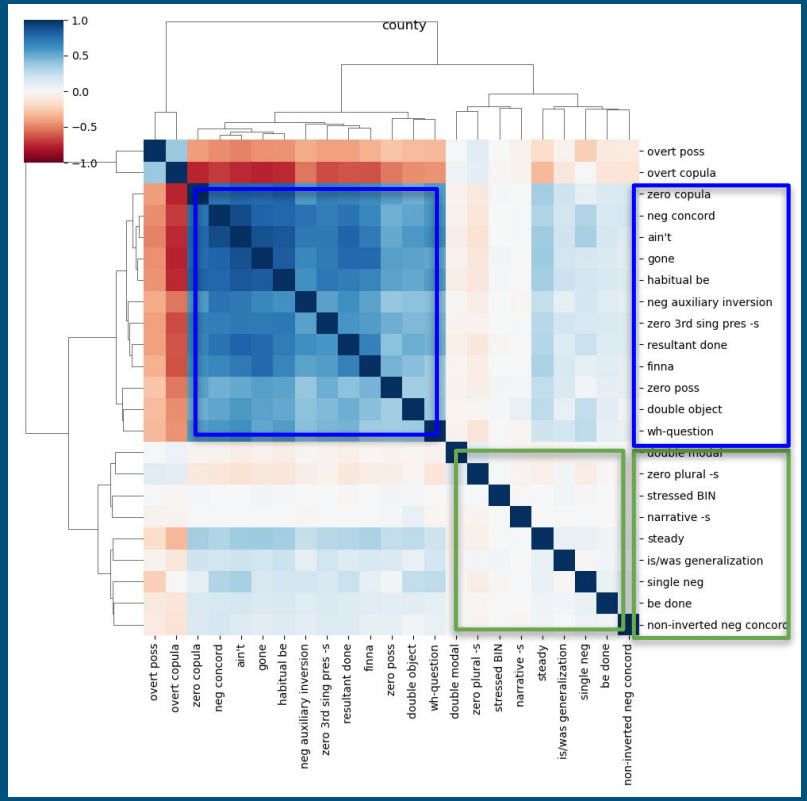
Preliminary analysis: feature co-occurrence

Group 1 - strong positive



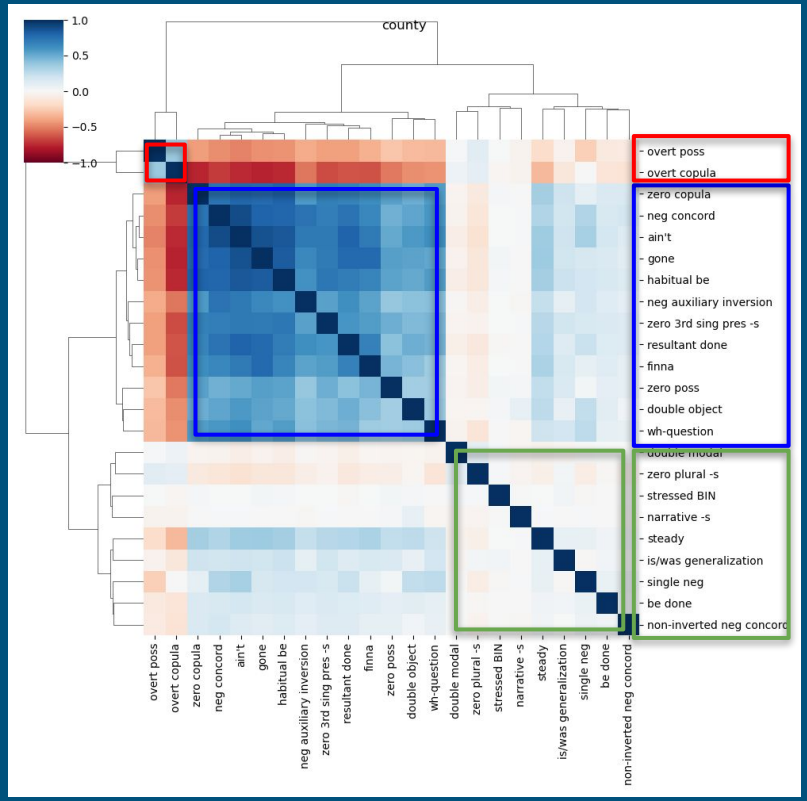
Preliminary analysis: feature co-occurrence

Group 1 - strong positive
Group 2 - mostly neutral



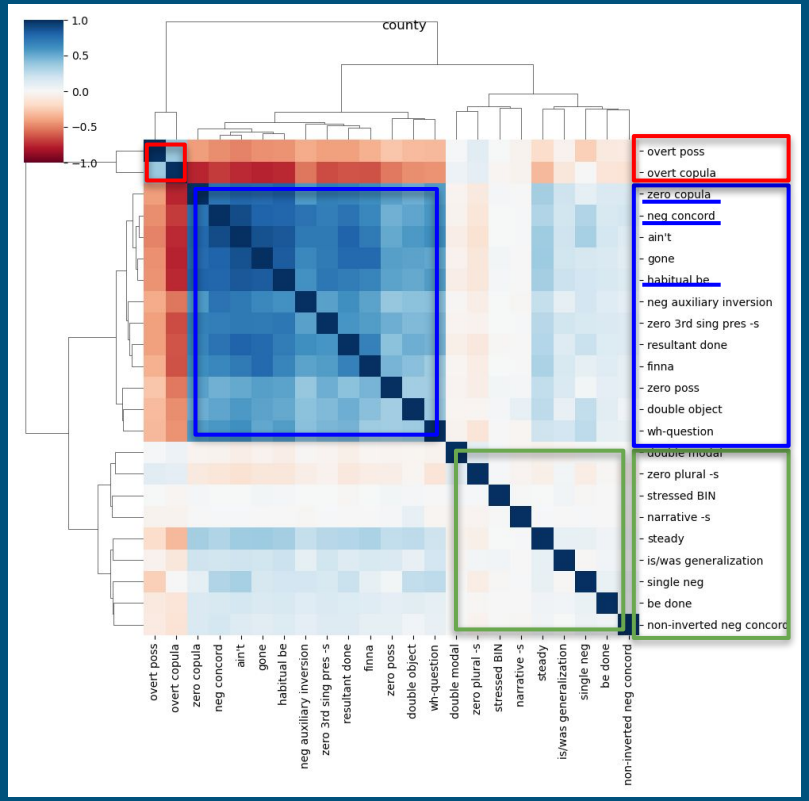
Preliminary analysis: feature co-occurrence

Group 1 - strong positive
Group 2 - mostly neutral
Group 3 - strong negative



Preliminary analysis: feature co-occurrence

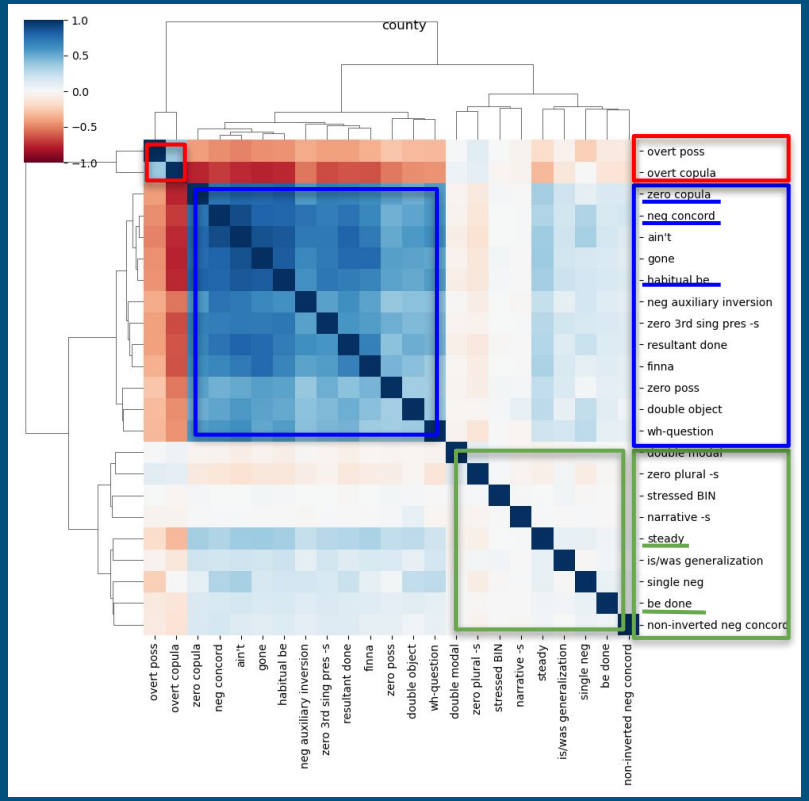
Group 1 - strong positive
Group 2 - mostly neutral
Group 3 - strong negative



zero copula
negative concord
habitual *be*

Preliminary analysis: feature co-occurrence

Group 1 - strong positive
Group 2 - mostly neutral
Group 3 - strong negative

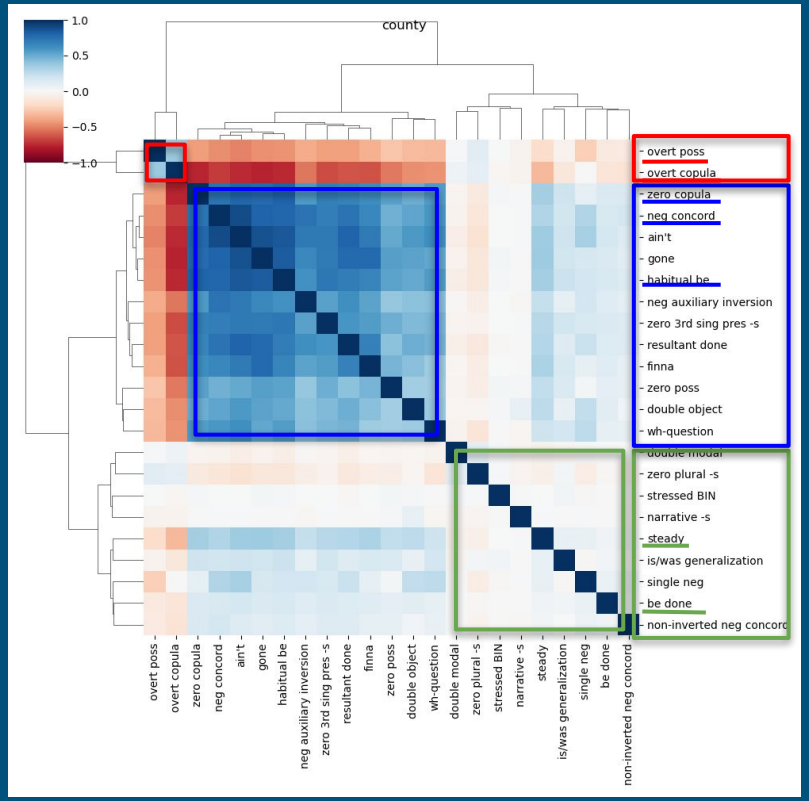


zero copula
negative concord
habitual *be*

steady
be done

Preliminary analysis: feature co-occurrence

Group 1 - strong positive
Group 2 - mostly neutral
Group 3 - strong negative



overt possessive
overt copula

zero copula
negative concord
habitual *be*

steady
be done

Future directions

- Identify systems of features/dialect groups
  - FDA or cluster analysis
  - Assign each county to a dialect group
- Map groups onto social factors
  - Are all counties in the dialect group also part of the social/regional group?

Future directions

- Identify systems of features/dialect groups
  - FDA or cluster analysis
  - Assign each county to a dialect group
- Map groups onto social factors
  - Are all counties in the dialect group also part of the social/regional group?
- Incorporating demographic information?

Future directions

- Identify systems of features/dialect groups
  - FDA or cluster analysis
  - Assign each county to a dialect group
- Map groups onto social factors
  - Are all counties in the dialect group also part of the social/regional group?
- Incorporating demographic information?

```
Relative incidence (feature) = (# tweets with feature / # total tweets)
```

Future directions

- Identify systems of features/dialect groups
  - FDA or cluster analysis
  - Assign each county to a dialect group
- Map groups onto social factors
  - Are all counties in the dialect group also part of the social/regional group?
- Incorporating demographic information?

```
Relative incidence (feature) = (# tweets with feature / # total tweets) *
  (African American blockgroup population / total blockgroup population)
```

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

Anissa Neal
anneal@linguist.umass.edu

Lisa Green
lgreen@linguist.umass.edu

Brendan O'Connor
brenocon@cs.umass.edu