Dialects in NLP

CS 685, Spring 2021

Advanced Topics in Natural Language Processing

http://brenocon.com/cs685

https://people.cs.umass.edu/~brenocon/cs685_s21/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

Presentations next Monday!

- One slide, no build/animations please - will send out a Google Slides URL
- 90 seconds per group

- Final report due **May 12** - last day of finals
  - Will also post optional "HW3" extra credit questions - due same day

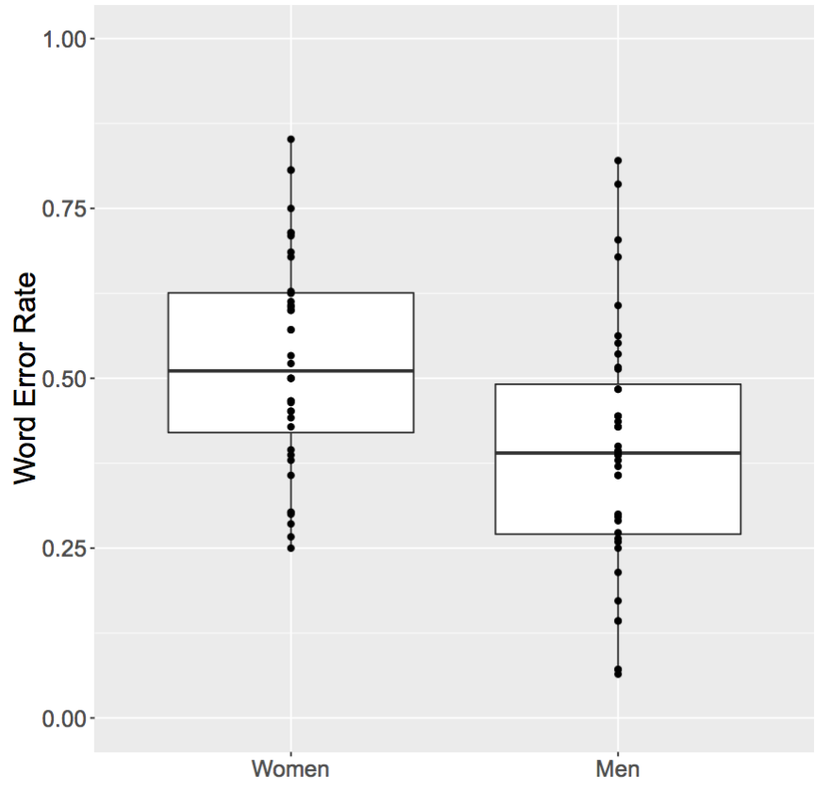
- Language is socially situated
- **<u>By</u>** and **<u>for</u>** communicators
  - Today - focus on *dialect*

- and **<u>about</u>** people

Author-conditional NLP disparity

- Language technologies analyze the linguistic behavior of people
- Language is affected by social context and attributes

Example: gender bias in YouTube autocaptions *[Tatman 2017]*



- What information can a user access?
- Whose voices are heard?

Variation in language

- Social factors drive language change and correlate with language varieties: by geography, ethnicity, gender, class...

- e.g. socioeconomic class: Labov (1966) finds more (r)-usage at more expensive New York department stores

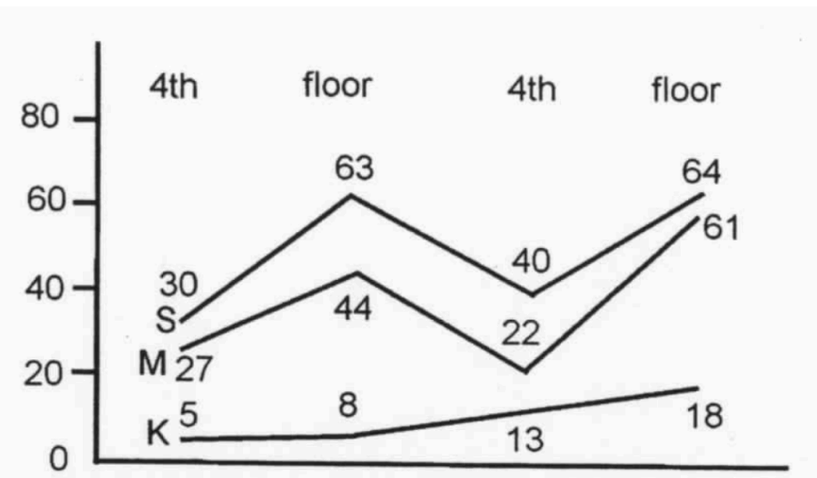


Figure 13.2: Percentage of all (r – 1) by store for four positions (S = Saks, M = Macy's, K = Kleins)

Variation in language

- Social factors drive language change and correlate with language varieties: by geography, ethnicity, gender, class...
  - Arbitrariness of language designations - "dialect is a language with an army and navy"
    - Cantonese, Moroccan Arabic ...
    - Swedish, Danish...
  - Distinct language varieties often associated with social groups and segregated communities
- We all know NLP domain adaptation is hard. How does this affect NLP performance?

Dialectal NLP

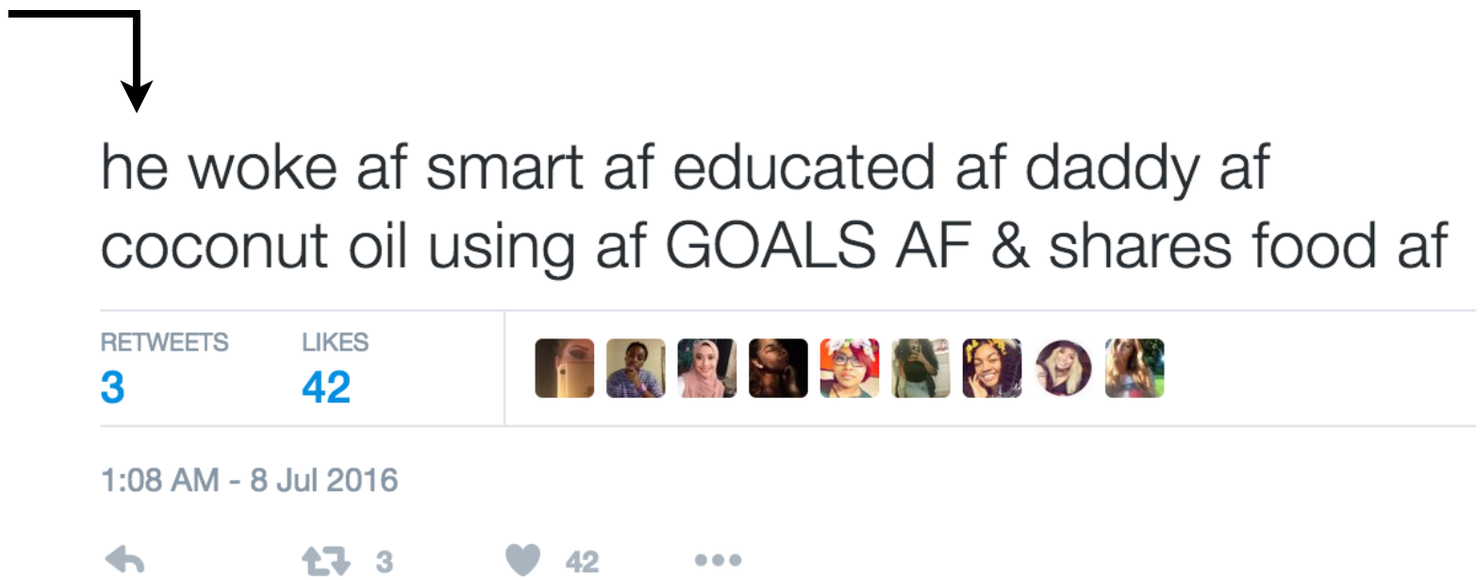
- 1. Identify the dialect & make a corpus
  - Through author-level metadata
  - Through in-text linguistic features

- 2. Evaluate NLP performance on the dialect

- 3. Adapt systems to work well on the dialect
  - Turns out to be tough!

Dialect in social media

SAE:
*he **is** woke af*

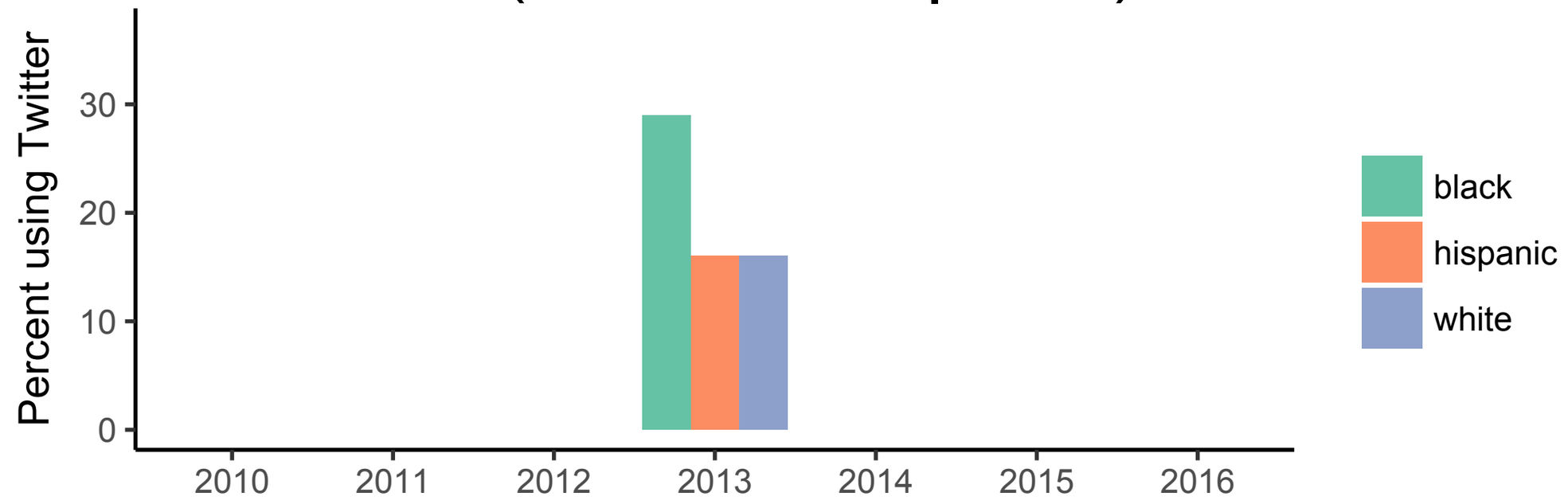


- **Why is social media different?**
  - Internet speech?
  - Pre-existing dialectal English?
    - Geographic patterns of word usage often reveal relationships to race, ethnicity etc.
    - African-American English in Twitter
      [Eisenstein 2013, Jorgensen et al. 2015, Jones 2015]

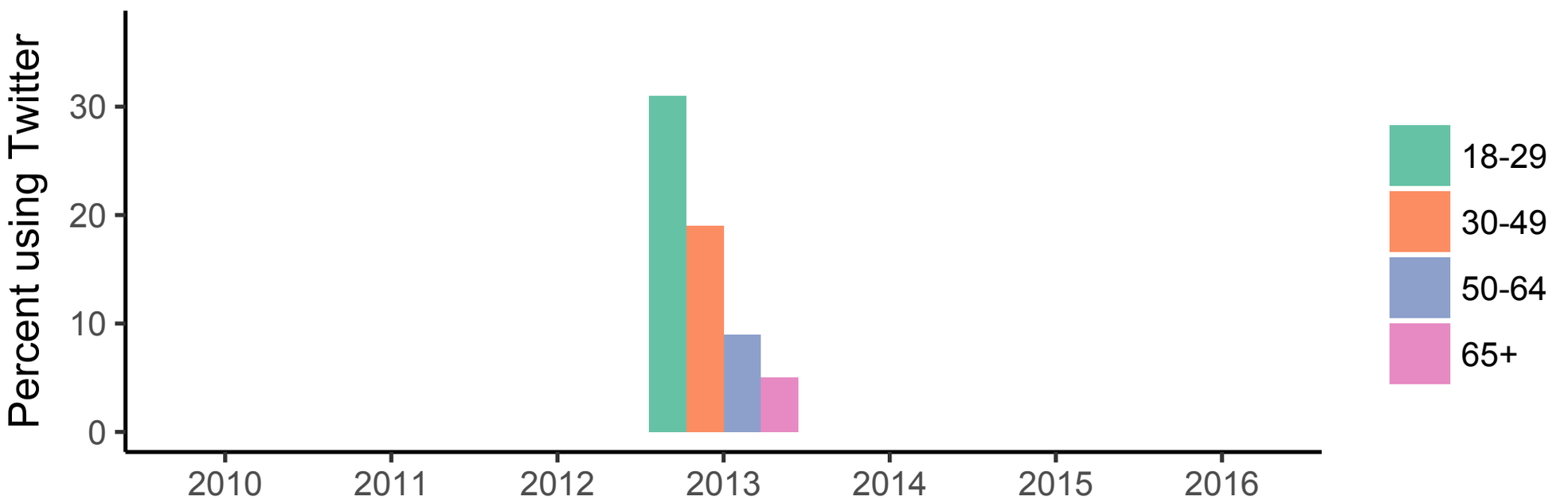
Youth, minorities on Twitter

[Pew Research]

P(use twitter | race)



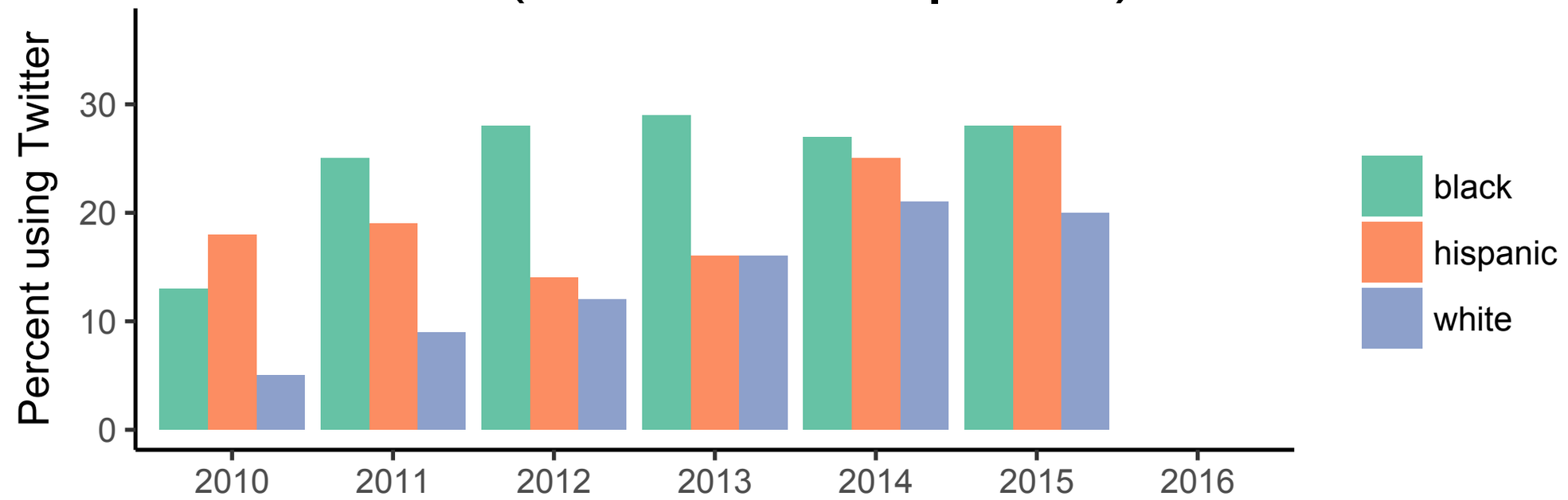
P(use twitter | age)



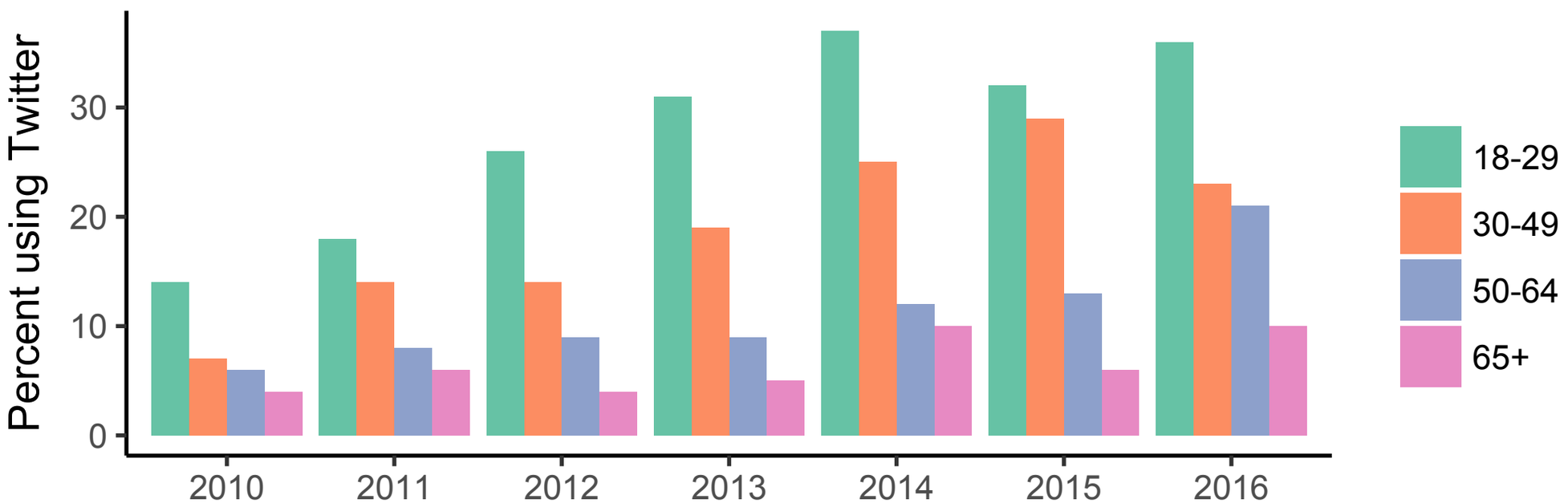
Youth, minorities on Twitter

[Pew Research]

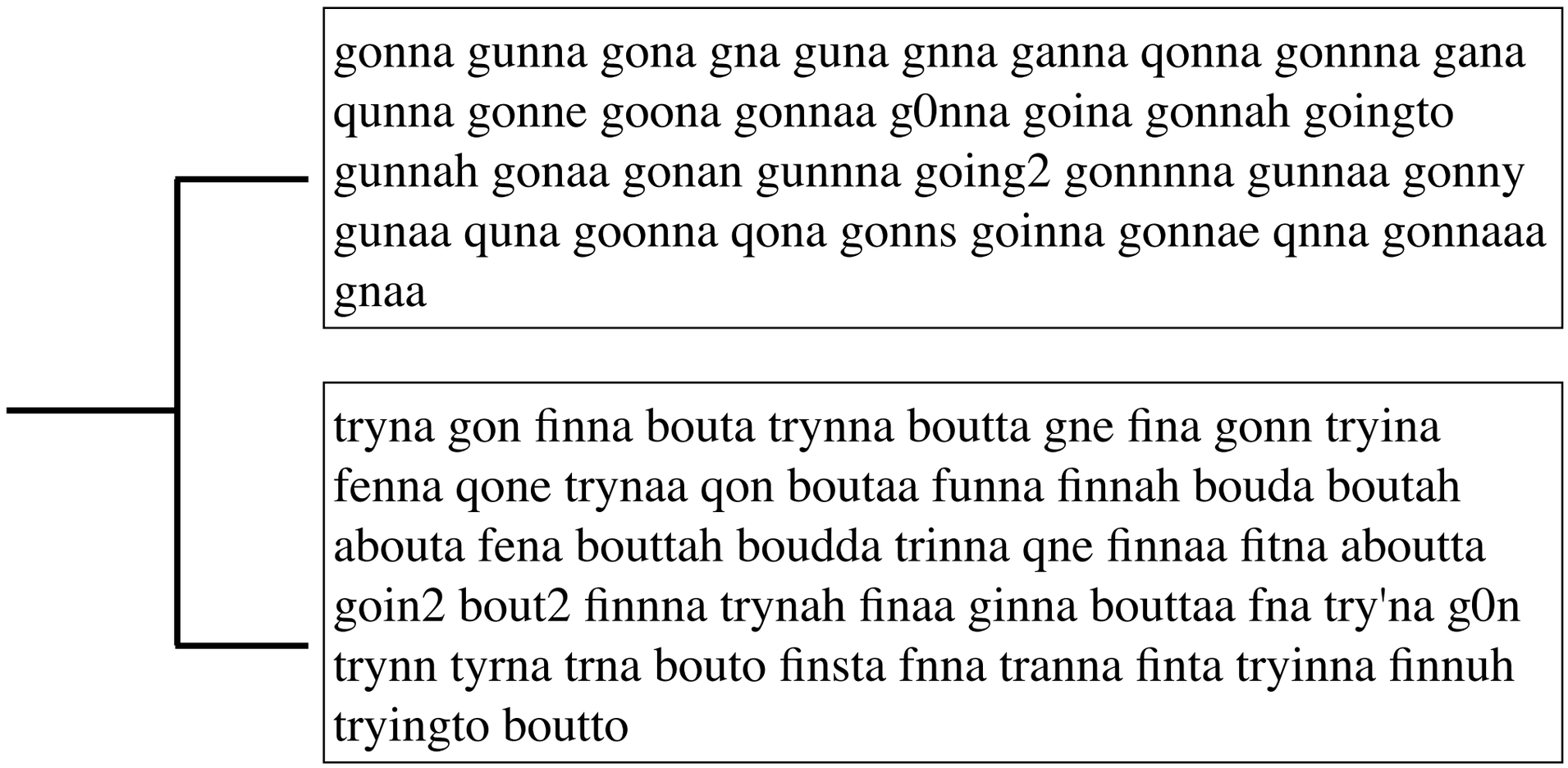
P(use twitter | race)



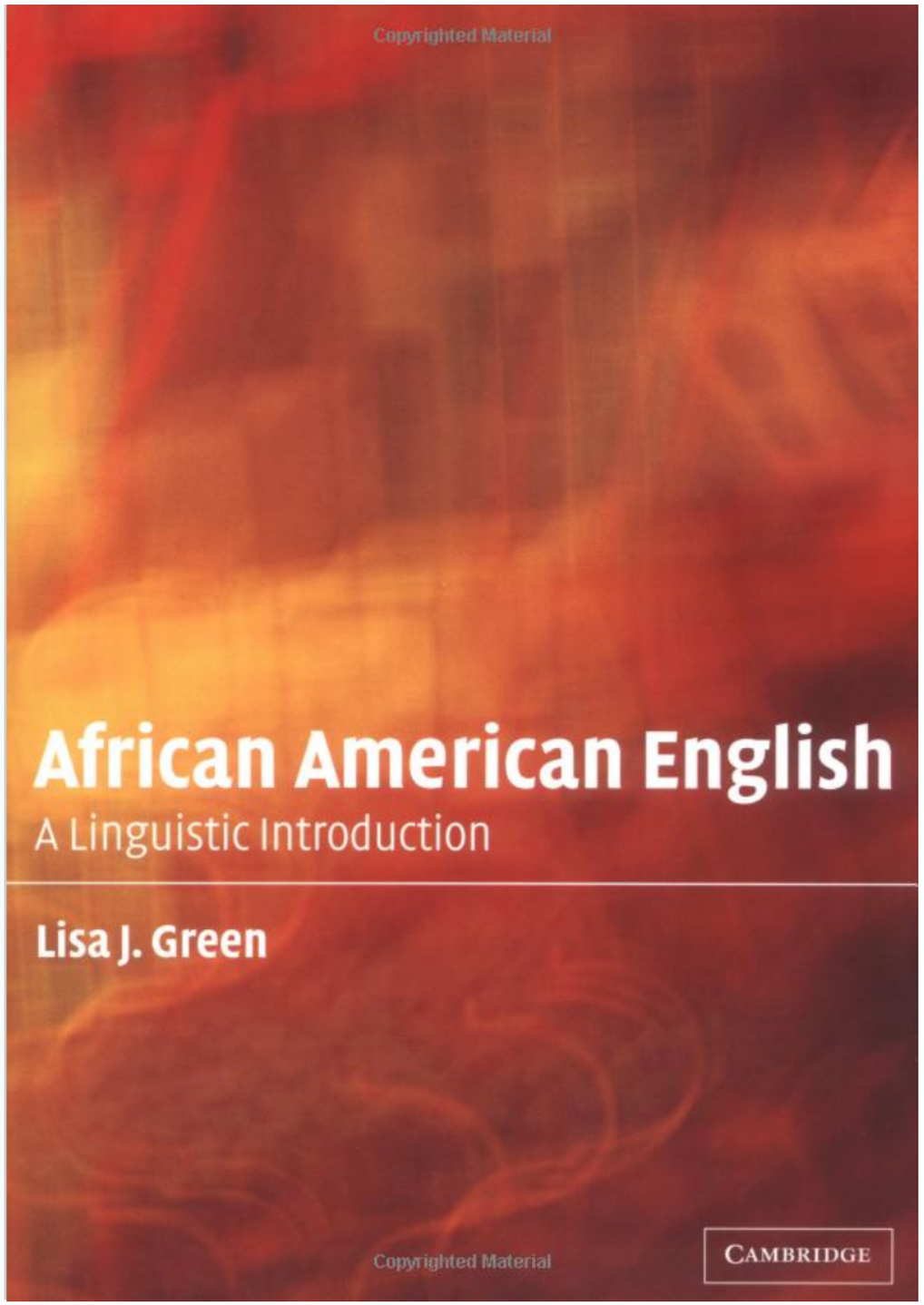
P(use twitter | age)



(Immediate?) future auxiliaries



- finna ~ “fixing to”
- tryna ~ “trying to”
- bouta ~ “about to”



2.4 Preverbal markers: *finna, steady, come*

Markers *finna*, *steady* and *come* have been identified in AAE, but they have not been analyzed to the extent that markers such as aspectual *be*, remote past *BIN* and *dən* have been analyzed. There are some descriptions of them in the literature, which will be cited in the summary of each preverbal marker. Also, note that lexical entries for *steady* and *come* are given in chapter 1.

Finna

Finna (including variants *fixina*, *fixna* and *fitna*) indicates that the event is imminent; it will happen in the immediate future. It precedes non-finite verbs, which are not marked for tense and agreement. Sentences in which this marker occurs are given below:

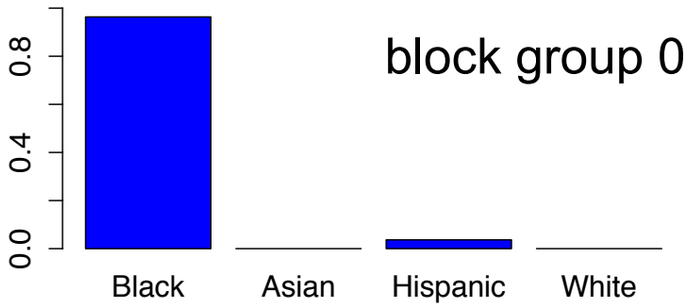
(77) a. I don't know about you, but I'm **finna leave**.
'I don't know about you, but I'm getting ready/about to leave'
b. Y'all **finna eat**?
'Are you getting ready/about to eat?'
c. She was **finna move** the mattress herself when I got there.
'She was getting ready/about to move the mattress when I got there'
d. Oh-oh they pulling they coats off. That mean they **fixna kill** us or something. (attested)

Associating geolocated tweets with demographics

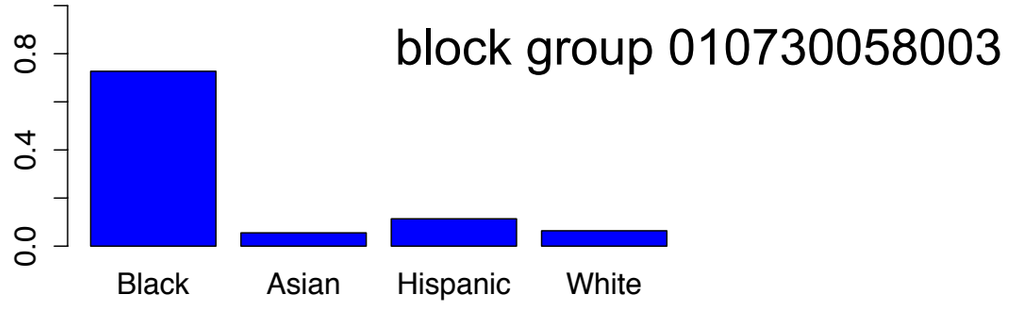
King Me
@tmac_datboi

Follow

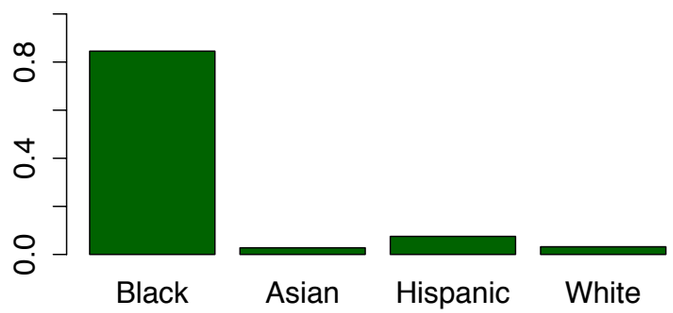
he woke af smart af educated af daddy af
coconut oil using af GOALS AF & shares food af



Bored af den my phone finna die!!!



π_{user} =



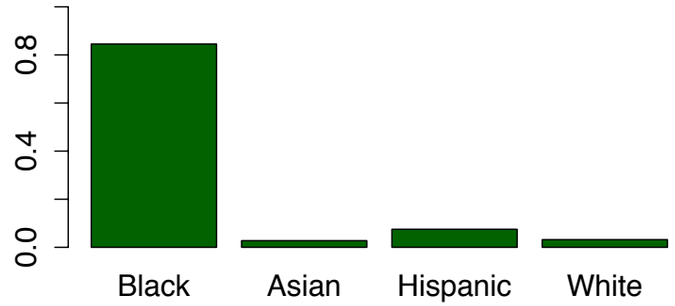
Mixed membership model

$\theta_{msg} \sim Dir(\alpha\pi)$, $z \sim \theta_{msg,}$ $w \sim \phi_z$

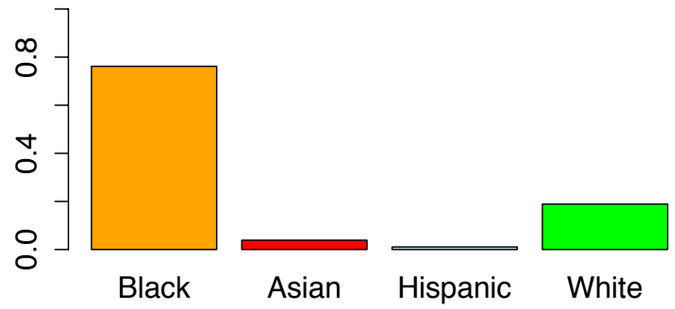
he woke af smart af educated af daddy af coconut oil using af GOALS AF & shares food af

Bored af den my phone finna die!!!

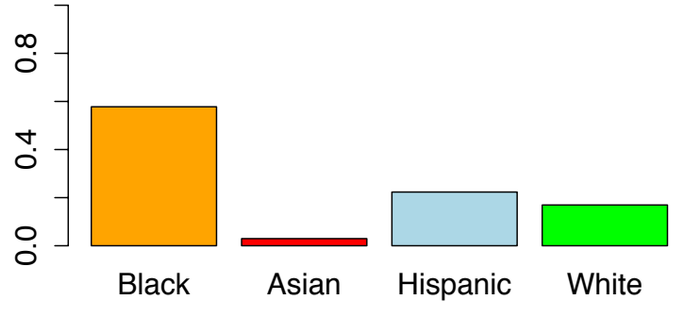
π_{user} =



θ_{msg} =



θ_{msg} =



Validation: Phonology

- For every word in vocabulary *w* and topic *k*, calculate

$$r_k(w) = \frac{p(w|z = k)}{p(w|z \neq k)}$$

- Calculate $r_{AA}(w)$ for 31 phonological variants illustrated through nonstandard spellings

- For 30/31 variants: $r \geq 1$

AAE	Ratio	SAE
sholl	1802.49	sure
iont	930.98	I don't
wea	870.45	where
talmbout	809.79	talking about
sumn	520.96	something

Model also has well-known syntactic phenomena in AAE (e.g. null copulas)

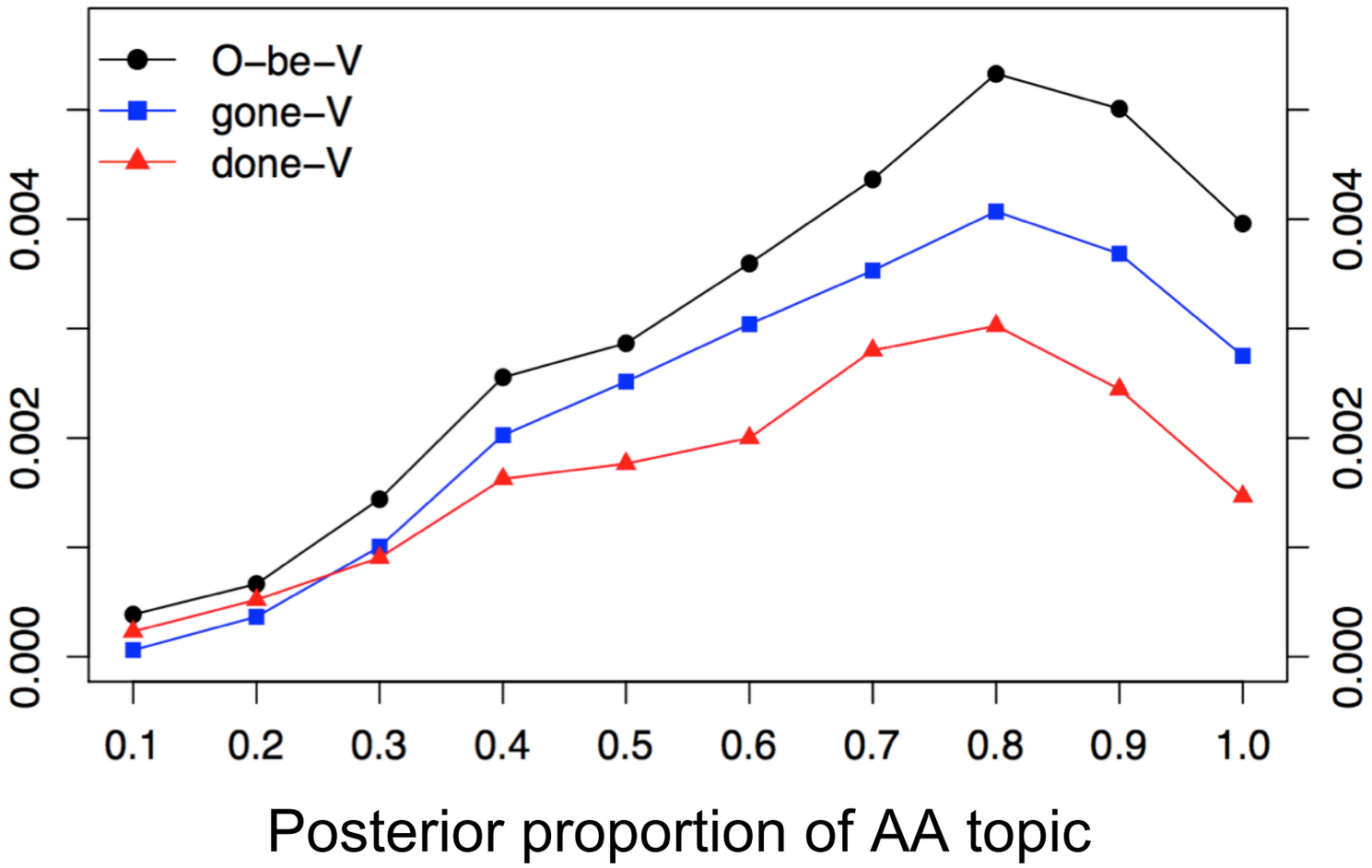
Validation: Syntax

- Select 3 well-known AAE verbal markers
- Search for sequences of unigrams and POS tags

Construction	**Example**
O-*be/b*-V	*I be tripping bruh*
gone/gne/gon-V	*Then she gon be single Af*
done/dne-V	*I done laughed so hard that I'm weak*

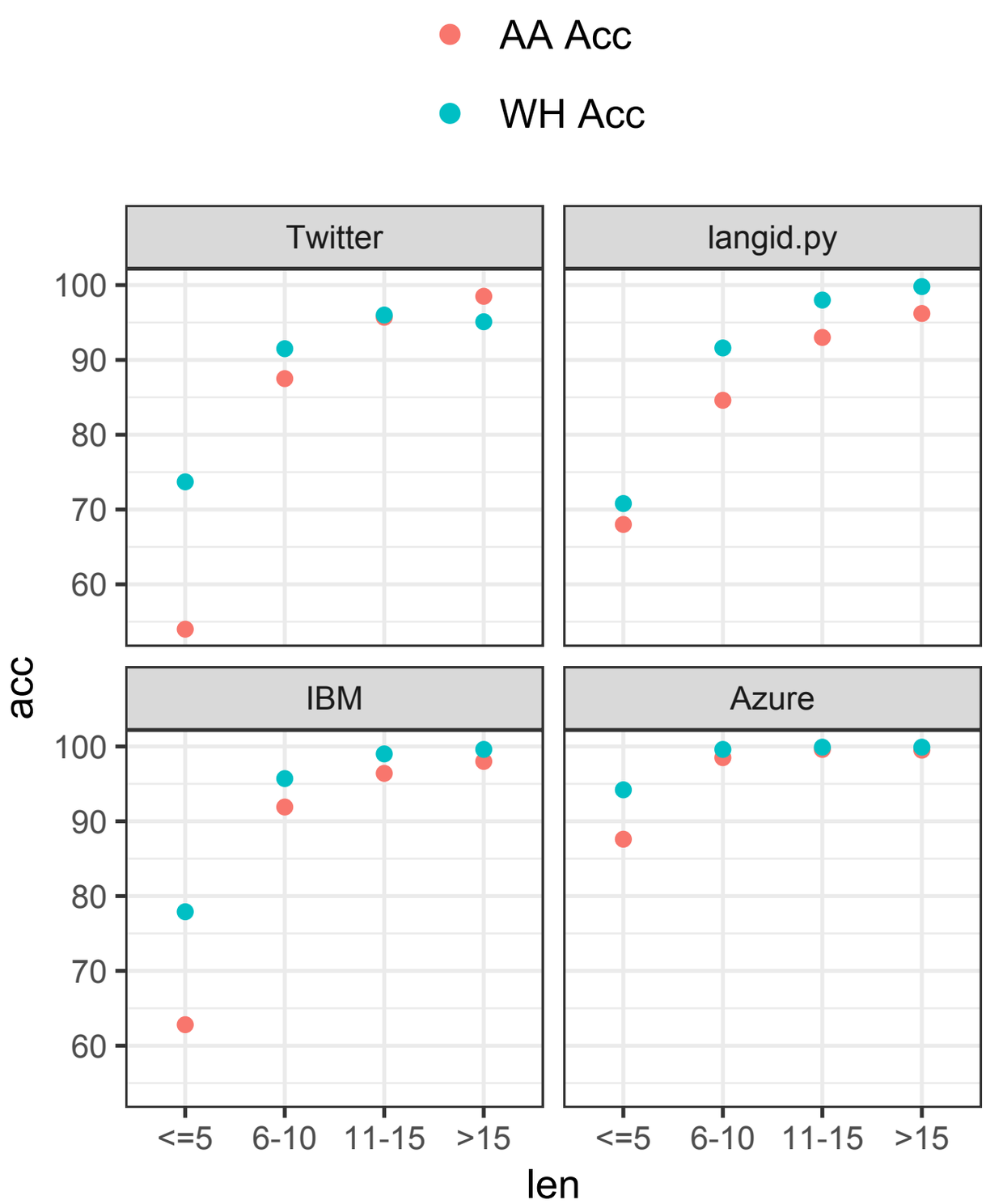
Validation: Syntax

Proportion of tweets with construction

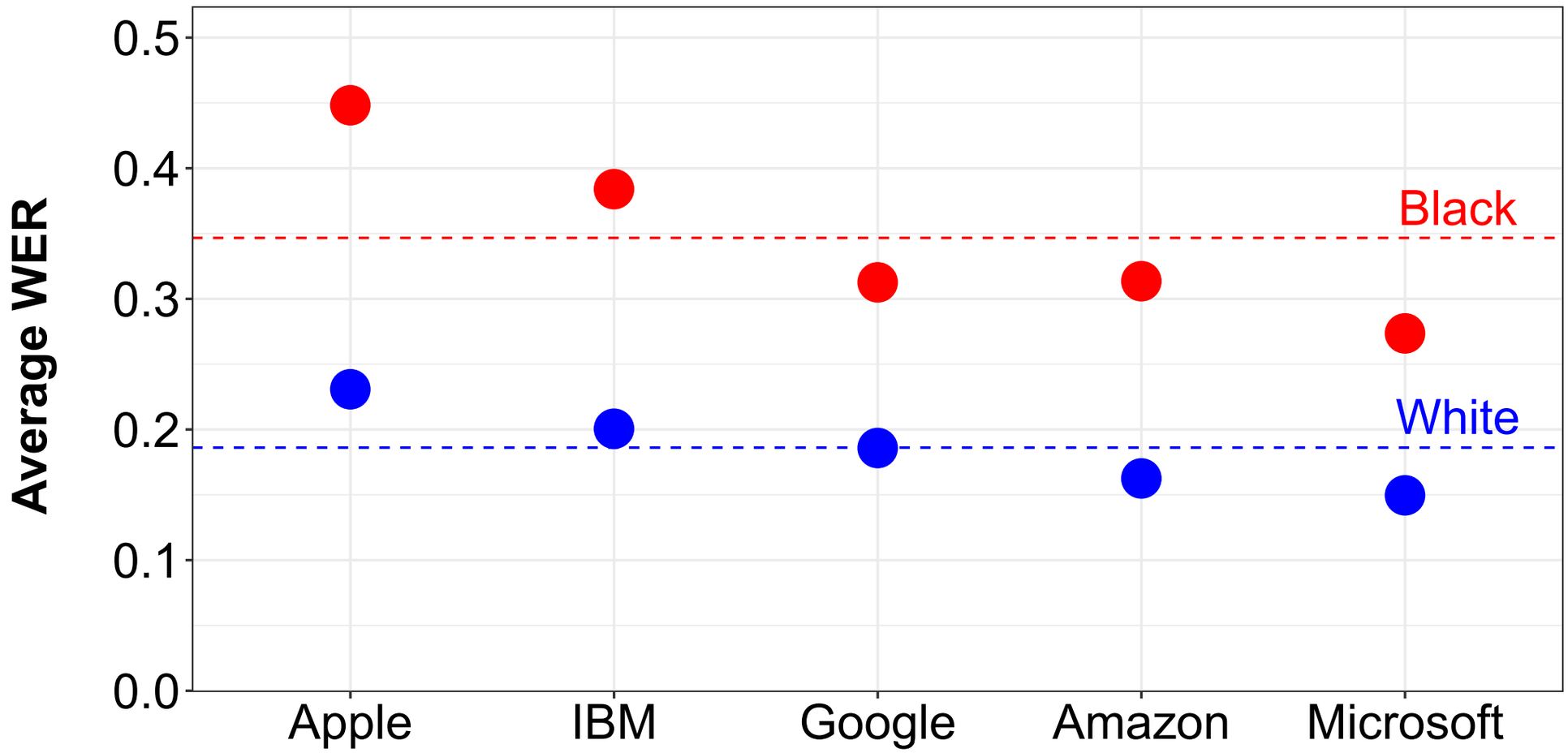


Racial disparity in accuracy

- p(correct | Wh) vs p(correct | AA)
- Assess disparity in
  - langid.py: popular open-source system *[Lui and Baldwin, 2012]*
  - Twitter (in metadata)
  - IBM, Microsoft



ASR and AAE

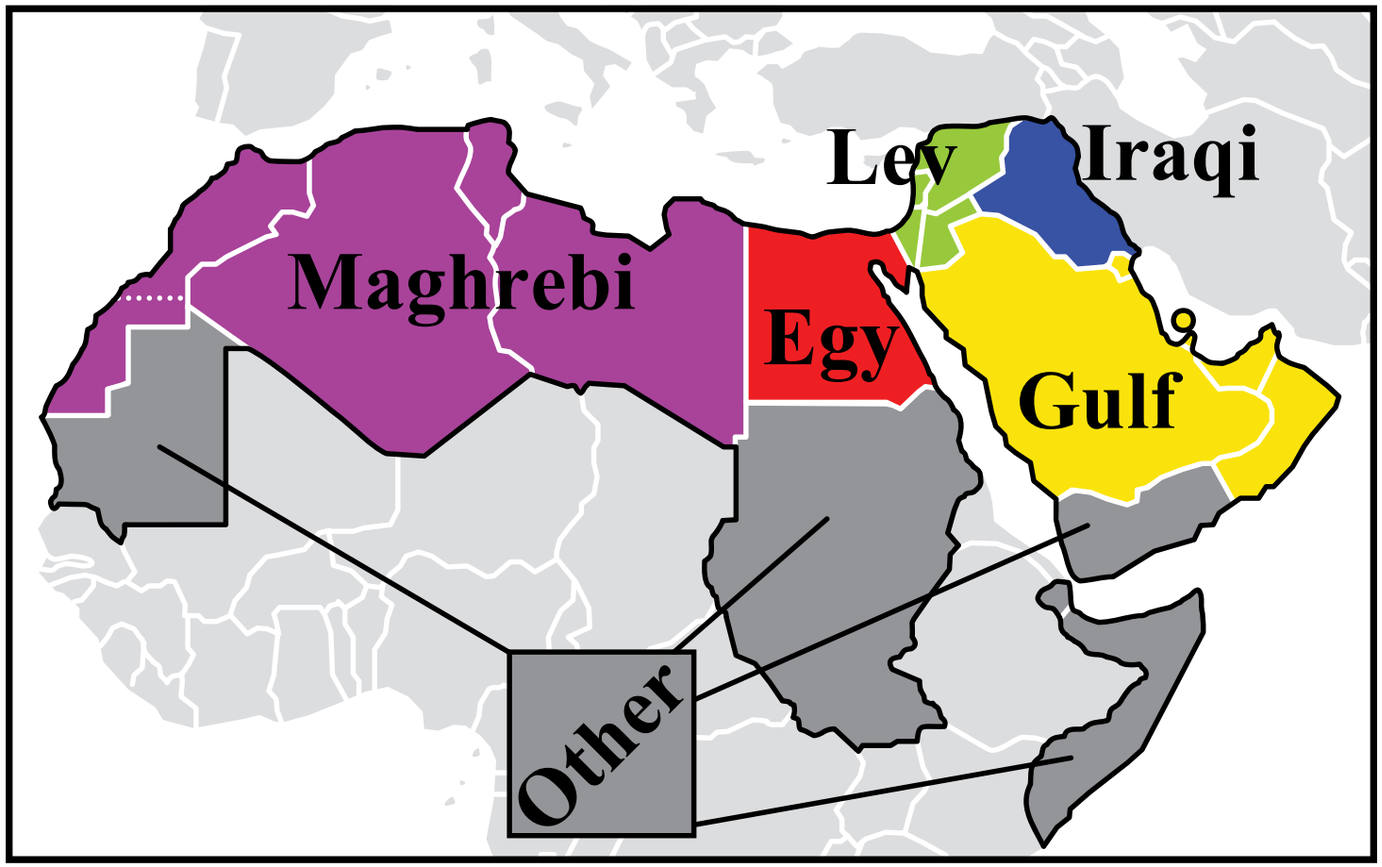


- From the Corpus of African-American Language (CORAAL): audio recordings of interviews

[Koenecke et al., 2020]

Direct identification of dialects

- Dialectal Arabic - annotated & classified from web text [*Zaidan and Callison-Burch, 2006*]



English	MSA	LEV	GLF	EGY
Book	*ktAb*	*ktAb*	*ktAb*	*ktAb*
Year	*snħ*	*snħ*	*snħ*	*snħ*
Money	*nqwd*	*mSAry*	*flws*	*flws*
Come on!	*hyA!*	*ylA!*	*ylA!*	*ylA!*
I want	*Aryd*	*bdy*	*Abγý*	*ςAyz*
Now	*AlĀn*	*hlq*	*AlHyn*	*dlwqt*
When?	*mtý?*	*Aymtý?*	*mtý?*	*Amtý?*
What?	*mAA?*	*Ayš?*	*wš?*	*Ayh?*
I drink	*Âšrb*	*bšrb*	*Ašrb*	*bšrb*
He drinks	*yšrb*	*bšrb*	*yšrb*	*byšrb*
We drink	*nšrb*	*bnšrb*	*nšrb*	*bnšrb*

Dialect ID from ling. features

- Linguistic knowledge-driven approach: identify sentences with particular *linguistic features*
- Supervise BERT fine-tuning with *minimal pairs*
- Application: identify Indian English

176. Deletion of copula *be*: before NPs
Feature area: Agreement
Typical example: He Ø a good teacher.

●	A - pervasive or obligatory	11
●	B - neither pervasive nor rare	13
●	C - extremely rare	7
●	D - absent	35
○	X - not applicable	7
○	? - no information	4

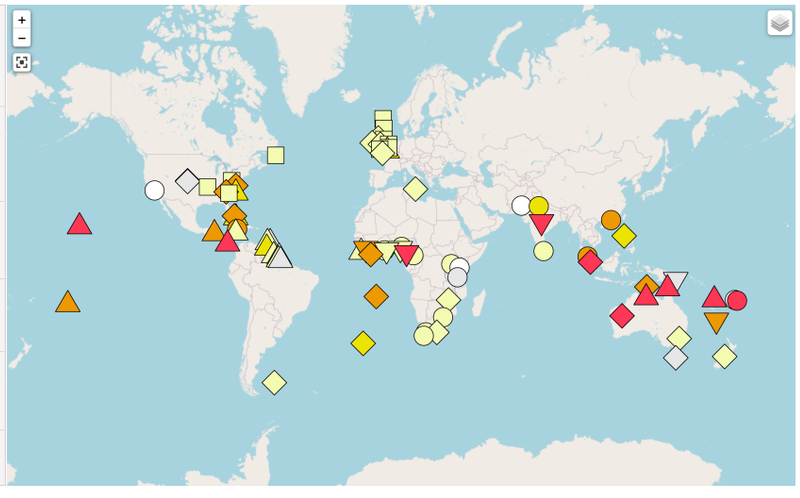


Figure 1: An example dialect feature from the Electronic World Atlas of Varieties of English (eWAVE).[1]

[Demszky et al., 2020]

Feature	**Example**	**Count of Instantiations**	
		Lange (2012)	Our data
ARTICLE OMISSION	*(the) chair is black*		59
DIRECT OBJECT PRO-DROP	*she doesn't like (it)*		14
FOCUS *itself*	*he is doing engineering in Delhi itself*	24	5
FOCUS *only*	*I was there yesterday only*	95	8
HABITUAL PROGRESSIVE	*always we are giving receipt*		2
STATIVE PROGRESSIVE	*he is having a television*		3
LACK OF INVERSION IN WH-QUESTIONS	*what you are doing?*		4
LACK OF AGREEMENT	*he do a lot of things*		23
LEFT DISLOCATION	*my father, he works for a solar company*	300	19
MASS NOUNS AS COUNT NOUNS	*all the musics are very good*		13
NON-INITIAL EXISTENTIAL	*every year inflation is there*	302	8
OBJECT FRONTING	*minimum one month you have to wait*	186	14
PP FRONTING WITH REDUCTION	*(on the) right side we can see a plate*		11
PREPOSITION OMISSION	*I went (to) another school*		17
INVERSION IN EMBEDDED CLAUSE	*I don't know what are they doing*		4
INVARIANT TAG *(isn't it, no, na)*	*the children are outside, isn't it?*	786	17
EXTRANEOUS ARTICLE	*she has a business experience*		25
GENERAL EXTENDER *and all*	*then she did her schooling and all*		7
COPULA OMISSION	*my parents (are) from Gujarat*	71	
RESUMPTIVE OBJECT PRONOUN	*my old life I want to spend it in India*	24	
RESUMPTIVE SUBJECT PRONOUN	*my brother, he lives in California*	287	
TOPICALIZED NON-ARGUMENT CONSTITUENT	*in those years I did not travel*	272	

Table 1: Features of Indian English used in our evaluations and their counts in the two datasets we study.

Minimal pairs

ARTICLE OMISSION: *chair is black* → *<u>the</u> chair is black*

FOCUS *only*: *I was there yesterday <u>only</u>* → *I was there just yesterday.*

NON-INITIAL EXISTENTIAL: *every year inflation <u>is there</u>* → *every year there is inflation.*

- Standard method of presentation in linguistics
- Demszky et al. use as supervision (one positive, one negative example) for BERT fine tuning to identify that particular feature

Dialect feature	**DAMTL**	**Multihead**
ARTICLE OMISSION	0.581	0.658
DIRECT OBJECT PRO-DROP	0.493	0.563
EXTRANEOUS ARTICLE	0.546	0.465
FOCUS *itself**	1.000	0.949
FOCUS *only**	0.998	0.775
HABITUAL PROGRESSIVE	0.439	0.718
INVARIANT TAG	0.984	0.901
INVERSION IN EMBEDDED CLAUSE	0.719	0.884
LACK OF AGREEMENT	0.543	0.674
LACK OF INVERSION IN WH-QUESTIONS	0.649	0.660
LEFT DISLOCATION	0.758	0.820
MASS NOUNS AS COUNT NOUNS	0.443	0.465
NON-INITIAL EXISTENTIAL*	0.897	0.885
OBJECT FRONTING	0.722	0.789
PREPOSITION OMISSION	0.500	0.648
PP FRONTING WITH REDUCTION	0.655	0.697
STATIVE PROGRESSIVE	0.645	0.789
GENERAL EXTENDER *and all*	0.994	0.991
Macro Average	0.698	0.741

Table 5: ROC-AUC results on the extended feature set, averaged across five random seeds. Because labeled

Social impact of NLP research

- Real-world language technology implementation
- Uses of language technologies
- Social biases in NLP models