Language from policy body camera footage shows racial disparities in officer respect

Voigt et al. 2017

Simone Zhang

Soc Stats Reading Group

March 13, 2019

Motivation

- There has been substantial public concern that the police treat black and white members of the community differently
- Past work on police-citizen interactions has relied on a) citizen recollections of past encounters or b) researcher observations of a limited set of interactions
- Body cams provide an opportunity to directly observe these interactions at scale

Data

- Transcripts of conversations between officers and black/white community members during traffic stops in Oakland, CA in April 2014
- 981 stops, 245 officers
- Transcripts divided up into utterances (a “turn” of one or more sentences)
- In total, there were 36,738 officer utterances

Descriptives

Total Officers		245
Race	White	102
	Black	39
	Asian	36
	Hispanic	57
	Other	11
Gender	M	224
	F	21
Mean Age		35.5 SD=8.2
Mean Years of Experience		7.1 SD=6.8
Mean Number of Stops in Dataset		4 SD=4.8

Community Member Race		Black	White
Total		682	299
Gender	M	463	177
	F	219	122
Mean Age		35.5 SD=13.6	38.4 SD=13.4
Stop Result	Arrest	40	1
	Citation	369	185
	Warning	273	113
Search Conducted	Yes	113	2
	No	569	297
Mean Stop Duration (Minutes)		12.6 SD=11.5	8.0 SD=5.1

Overview of the paper's approach

1. Draw a sample of officer utterances
2. Hire human annotators to rate the tone of the utterances in the sample
3. Build a model that predicts human ratings of tone
4. Apply model from previous step to estimate tone of all officer utterances
5. Test whether officers speak to black community members less respectfully

Outline

1. Measuring tone (Study 1)

2. Modeling tone (Study 2)
   - Extracting features from text
   - NLP tools
   - Feature selection and modeling
   - Validation

3. Testing for racial difference in tone (Study 3)
   - Main analysis
   - Linguistic classification accuracy of race

Rating task

- Sampled 414 unique officer utterances (about 1%)
  - Limited to utterances where 1) least 15 words were spoken between the two speakers, and 2) at least five words were spoken by the officer.
- Each utterance was rated by 10 different human coders
- Human coders were presented with
  - What the officer said
  - What the driver said right before that
- Human coders rated what the officer said on a scale from 1-4 on five "folk notions related to respect and officer treatment":
  1. Disrespectful - respectful
  2. Impolite - polite
  3. Judgmental - impartial
  4. Unfriendly - friendly
  5. Informal - formal

Inter-rater agreement

The authors present Cronbach's α by batch:

Batch	Formal	Friendly	Impartial	Polite	Respectful
1	0.82	0.86	0.84	0.86	0.83
2	0.88	0.89	0.86	0.86	0.87
3	0.80	0.87	0.73	0.84	0.78
4	0.85	0.91	0.79	0.88	0.87
5	0.77	0.89	0.81	0.87	0.87
6	0.91	0.82	0.81	0.87	0.86
7	0.85	0.86	0.84	0.84	0.84

Table 4: Annotator consistency (Cronbach's α) across batches and dimension for the utterance-level thin-slice judgments in Study 1.

Cronbach's α reflects internal consistency

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_x^2}\right) \tag{1}$$

where k is number of coders, $\sigma_{y_i}^2$ is variance of the sum of the coder ratings, $\sigma_{y_i}^2$ is variance of individual i's ratings

The authors say:

> *These results demonstrate the transcribed language provides a sufficient and consensual signal of officer communication, enough to gain a picture of the dynamics of an interaction at a given point in time.*

Under what conditions might we not fully be convinced this is the case?

- If we believe that different people perceive tone differently and the raters are non-representative in consequential ways
  - 70 raters (56% female, median age 25).
- If the community member utterances provide cues about the speaker's race, affecting ratings of officer utterances
- If a large component of tone in spoken conversations is lost or distorted when presented on paper as text.

Principal Component Analysis

Final rating for each utterance along each dimension was the average across the 10 raters.

Authors then used PCA to decompose the ratings into two underlying components:

	PC1: Respect	PC2: Formality
Formal	0.272	0.913
Friendly	0.464	−0.388
Impartial	0.502	−0.113
Polite	0.487	−0.047
Respectful	0.471	0.026
% of Variance Explained	71.3%	21.9%

Explained 93.2% of variance in ratings overall.

Outline

1. Measuring tone (Study 1)

2. Modeling tone (Study 2)
   - Extracting features from text
   - NLP tools
   - Feature selection and modeling
   - Validation

3. Testing for racial difference in tone (Study 3)
   - Main analysis
   - Linguistic classification accuracy of race

The features they extracted

Feature Name	Implementation	Source
Adverbial "Just"	"Just" occurs in a dependency arc as the head of an `advmod` relation	
Apologizing	Lexicon: `"sorry"`, `"oops"`, `"woops"`, `"excuse me"`, `"forgive me"`, `"apologies"`, `"apologize"`, `"my bad"`, `"my fault"`	[4]
Ask for Agency	Lexicon: `"do me a favor"`, `"let me"`, `"allow me"`, `"can i"`, `"should i"`, `"may i"`, `"might i"`, `"could i"`	[4]
Bald Command	The first word in a sentence is a bare verb with part-of-speech tag VB (`"look"`, `"give"`, `"wait"` etc.) but is not one of `"be"`, `"do"`, `"have"`, `"thank"`, `"please"`, `"hang"`.	
Colloquialism	Regular expression capturing `"y'all"`, `"ain't"` and words ending in `"in'"` such as `"walkin'"`, `"talkin'"`, etc., as marked by transcribers	
Conditional	Lexicon: `"if"`	
Disfluency	Word fragment ("Well I thi-") as indicated by transcribers	[5, 6]
Filled Pauses	Lexicon: `"um"`, `"uh"`	[7, 8]
First Names	Top 1000 most common first names from the 1990 US Census, where first letter is capitalized in transcript	[9, 10][1]
Formal Titles	Lexicon: `"sir"`, `"ma'am"`, `"maam"`, `"mister"`, `"mr*"`, `"ms*"`, `"madam"`, `"miss"`, `"gentleman"`, `"lady"`	[9, 10]
For Me	Lexicon: `"for me"`	
For You	Lexicon: `"for you"`	
Give Agency	Lexicon: `"let you"`, `"allow you"`, `"you can"`, `"you may"`, `"you could"`	[4]
Gratitude	Lexicon: `"thank"`, `"thanks"`, `"appreciate"`	[4]
Goodbye	Lexicon: `"goodbye"`, `"bye"`, `"see you later"`	

The features they extracted

<table>
<tr><td>Hands on the Wheel</td><td>Regular expression capturing cases like "keep your hands on the wheel" and "leave your hands where I can see them": <code>"hands? ([.,?!:;]+)?(wheel|see)"</code></td><td></td></tr>
<tr><td>Hedges</td><td>All words in the "Tentat" LIWC lexicon</td><td>[11]</td></tr>
<tr><td>Impersonal Pronoun</td><td>All words in the "Ippron" LIWC lexicon</td><td>[4, 11]</td></tr>
<tr><td>Informal Titles</td><td>Lexicon: <code>"dude*"</code>, <code>"bro*"</code>, <code>"boss"</code>, <code>"bud"</code>, <code>"buddy"</code>, <code>"champ"</code>, <code>"man"</code>, <code>"guy*"</code>, <code>"guy"</code>, <code>"brotha"</code>, <code>"sista"</code>, <code>"son"</code>, <code>"sonny"</code>, <code>"chief"</code></td><td>[9, 10, 12]</td></tr>
<tr><td>Introductions</td><td>Regular expression capturing cases like "I'm Officer [name] from the OPD" and "How's it going?": <code>"((i|my name).+officer | officer.+(oakland|opd))|((hi|hello|hey|good afternoon|good morning|good evening|how are you doing|how 's it going))"</code></td><td>[4]</td></tr>
<tr><td>Last Names</td><td>Top 5000 most common last names from the 1990 US Census, where first letter is capitalized in transcript</td><td>[9, 10][2]</td></tr>
<tr><td>Linguistic Negation</td><td>All words in the "Negate" LIWC lexicon</td><td>[11]</td></tr>
<tr><td>Negative Words</td><td>All words in the "Negativ" category in the Harvard General Inquierer, matching on word lemmas</td><td>[4, 13]</td></tr>
<tr><td>Positive Words</td><td>All words in the "Positiv" category in the Harvard General Inquierer, matching on word lemmas</td><td>[4, 13]</td></tr>
</table>

The features they extracted

<table>
<tr><td>Please</td><td>Lexicon: "please"</td><td>[4]</td></tr>
<tr><td>Questions</td><td>Occurrence of a question mark</td><td></td></tr>
<tr><td>Reassurance</td><td>Lexicon: "'s okay", "n't worry", "no big deal", "no problem", "no worries", "'s fine", "you 're good", "is fine", "is okay"</td><td></td></tr>
<tr><td>Safety</td><td>Regular expression for all words beginning with the prefix "safe", such as "safe", "safety", "safely"</td><td></td></tr>
<tr><td>Swear Words</td><td>All words in the "Swear" LIWC lexicon</td><td>[11]</td></tr>
<tr><td>Tag Question</td><td>Regular expression capturing cases like "..., right?" and "..., don't you?": ", (((all right|right|okay|yeah|please|you know)(sir| ma'am| miss| son)?)|((are|is|do|can|have|will|won't) (n't)?(i|me|she|us|we|you|he|they|them))) [?]"</td><td>[14, 15]</td></tr>
<tr><td>The Reason for the Stop</td><td>Lexicon: "reason", "stop* you", "pull* you", "why i", "why we", "explain", "so you understand"</td><td></td></tr>
<tr><td>Time Minimizing</td><td>Regular expression capturing cases like "in a minute" and "let's get this done quick": "(a|one|a few) (minute|min|second|sec|moment)s?|this[␣,?!]+quick|right back"</td><td></td></tr>
</table>

NLP tools in R

General solutions
For tokenization, part of speech tagging, named entity recognition, entity linking, sentiment analysis, dependency parsing, coreference resolution, and word embeddings:

- `openNLP`: provides wrapper for openNLP (Java)
- `cleanNLP`: provides wrapper for spaCy (Python), Stanford CoreNLP (Java), udpipe (C++)

More specific to markers of politeness

- `politeness`: based on past work identifying linguistic markers of politeness

Feature selection

Used simple linear regression and stepwise feature selection by R^2.

- Authors state that they also tried modeling using lasso, support vector regression, and random forest with the same set of features but performance was no better

Outcome variables: respect and formality

Independent variables: log counts of linguistic features at utterance level.

3.2 Full Regression Model Output

	Respect			*Formality*		
	β	CI	p	β	CI	p
Fixed Parts						
(Intercept)	-0.18	-0.36 – 0.00	.052	0.26	0.07 – 0.45	**.008**
Adverbial "Just"	0.24	-0.07 – 0.53	.118			
Apologizing	1.34	0.15 – 2.52	**.027**	-1.56	-2.80 – -0.32	**.014**
Ask for Agency	-0.34	-0.90 – 0.22	.230	0.37	-0.23 – 0.96	.225
Bald Commands				-0.25	-0.68 – 0.18	.255
Colloquialism				-1.10	-1.97 – -0.23	**.013**
Conditional				-0.27	-0.74 – 0.21	.271
Disfluency	-0.36	-0.63 – -0.09	.009			
Filled Pauses (Um/Uh)	0.37	0.14 – 0.60	**.002**	-0.40	-0.64 – -0.16	**.001**
First Names	-0.88	-1.66– -0.11	**.026**			
Formal Titles	0.73	0.20 – 1.26	**.007**	0.96	0.43 – 1.49	$<$**.001**
For Me	0.56	-0.08 – 1.21	.086			
For You	1.08	-0.70 – 2.87	.234	-1.26	-3.10 – 0.58	.178
Give Agency	0.39	0.01 – 0.78	**.047**	0.40	-0.02 – 0.82	.063
Gratitude	1.04	0.44 – 1.64	$<$**.001**			
Hands on the Wheel	-1.09	-2.27 – 0.07	.065	1.33	0.10 – 2.55	**.034**
Hedges	0.18	0.00 – 0.37	.053			
Impersonal Pronouns				-0.10	-0.27 – 0.07	.269
Informal Titles	-0.65	-1.03 – -0.28	$<$**.001**	-1.06	-1.45 – -0.68	$<$**.001**
Introductions	0.18	-0.12 – 0.48	.235			
Last Names	0.75	0.39 – 1.12	$<$**.001**	0.26	-0.10 – 0.62	.156
Linguistic Negation	-0.22	-0.43 – -0.03	**.027**	0.22	0.01 – 0.43	**.045**
Negative Words	-0.24	-0.40 – -0.07	**.005**	-0.17	-0.34 – 0.01	.056
Positive Words	0.20	0.03 – 0.37	**.020**	-0.16	-0.32 – 0.00	.056
Questions	-0.20	-0.43 – 0.02	.075	0.26	0.02 – 0.49	**.031**
Reassurance	1.04	0.34 – 1.74	**.004**	-0.73	-1.46 – 0.00	**..049**
Safety	0.54	0.06 – 1.02	**.027**			
The Reason for the Stop				0.41	0.08 – 0.75	**.015**
Time Minimizing				-0.66	-1.31 – 0.00	**.049**
Observations		414			414	
R^2 / Ω_0^2		.298 / .258			.229 / .190	

Table 9: Linear regression outputs, with stepwise feature selection by R^2, for all annotated utterances with *Respect* and *Formality* (PC1 and PC2) as dependent variables and utterance-level log counts of linguistic features as independent variables. The swear words, please, goodbye, and tag question features were selected out in both models.

Assigning respect scores

Example	Respect Score
First Name · Ask For Agency · Questions [name], can I see that driver's license again? It- it's showing suspended. Is that- that's you? Disfluency · Negative Word · Disfluency	-1.07
Informal Title · Ask For Agency · Adverbial "Just" All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick. "Hands On The Wheel"	-0.51
Apology · Introduction · Last Name Sorry to stop you. My name's Officer [name] with the Police Department.	0.84
Formal Title · Safety · Please There you go, ma'am. Drive safe, please.	1.21
Adverbial "Just" · Filled Pause · Reassurance It just says that, uh, you've fixed it. No problem. Thank you very much, sir.	2.07

Validation

We are interested in whether the model does a good job of predicting how people actually rate.

How do the predicted ratings compare to actual human ratings?

Assessing performance

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

where

- i indexes an officer utterance
- y_i is human rating for utterance i
- $\hat{y}_i$ is predicted rating for utterance i
- n is the number of utterances (n = 414)

Assessing performance

RMSE for Respect: 0.84; **RMSE for Formality**: 0.88

How to assess if this is good? What the authors do:

- Benchmark in comparison to RMSE *across human coders*
- Treat the average rating as a gold standard

	Mean	Median	Max	Min
Respect	0.842	0.826	1.677	0.497
Formality	0.764	0.718	1.703	0.518

Table 10: Human RMSE scores for *Respect* and *Formality* across annotators relative to other annotators.

Outline

Main question

From pg. 6523:

> *Controlling for contextual factors of the interaction, is officers' language more respectful when speaking to white as opposed to black community members?*

Strategy

Apply models from previous stage to rate all utterances for Respect and Formality.

Estimate linear mixed effect models:

- Outcome variables: Respect and Formality
- Covariates:
  - Community member race, age, and gender
  - Officer race
  - Whether a search was conducted
  - The result of the stop (warning, citation, arrest)
- Random intercepts for interactions nested within officers

Results

"Controlling for these contextual factors, utterances spoken by officers to white community members score higher in Respect."

	Respect			*Formality*		
	β	CI	p	β	CI	p
Fixed Parts						
Arrest Occurred	0.00	-0.03 – 0.03	.933	0.01	-0.02 – 0.04	.528
Citation Issued	0.04	0.02 – 0.06	<**.001**	0.01	-0.01 – 0.03	.209
Search Conducted	-0.08	-0.11 – -0.05	<**.001**	0.00	-0.03 – 0.02	.848
Age	0.07	0.05 – 0.09	<**.001**	0.05	0.03 – 0.07	<**.001**
Gender (F)	0.02	0.00 – 0.04	.062	0.02	0.00 – 0.04	**.025**
Race (W)	0.05	0.03 – 0.08	<**.001**	-0.01	-0.04 – 0.01	.236
Officer Race (B)	0.00	-0.03 – 0.04	.884	0.00	-0.03 – 0.03	.987
Officer Race (O)	0.00	-0.04 – 0.03	.809	0.00	-0.03 – 0.02	.783
Officer Race (B) : Race (W)	-0.01	-0.03 – 0.02	.583	0.01	-0.01 – 0.03	.188
Officer Race (O) : Race (W)	-0.01	-0.03 – 0.02	.486	0.00	-0.02 – 0.02	.928
Random Parts						
σ^2		0.918			0.954	
$\tau_{00,\text{Stop:Officer}}$		0.045			0.029	
$\tau_{00,\text{Officer}}$		0.029			0.015	
$N_{\text{Stop:Officer}}$		981			981	
N_{Officer}		245			245	
$ICC_{\text{Stop:Officer}}$		0.045			0.029	
ICC_{Officer}		0.029			0.015	
Observations		36738			36738	
R^2 / Ω_0^2		.100 / .097			.064 / .059	

Over time

To see how scores change over the course of an interaction, added a random slope of utterance position (where in conversation the utterance happened, scale 0 - 1)

	Respect			*Formality*		
	b	CI	p	*b*	CI	p
Fixed Parts						
Intercept	0.05	0.01 – 0.08	**<.001**	0.00	-0.02 – 0.02	.72
Race (W)	0.20	0.15 – 0.25	**<.001**	0.00	-0.04 – 0.04	.88
Utterance Position (mean-centered)	0.24	0.19 – 0.29	**<.001**	-0.48	-0.52 – -0.45	**<.001**
Utterance Position: Race (W)	0.20	0.10 – 0.31	**<.001**	-0.18	-0.27 – -0.10	**<.001**
Random Parts						
σ^2		0.90			0.93	
$\tau_{00,\text{Stop}}$		0.09			0.05	
$\tau_{11,\text{Utterance Position}}$		0.23				
$cor_{\tau_{00},\tau_{11}}$		-0.24				
N_{Stop}		981			981	
ICC_{Stop}		0.09			0.05	
Observations		36,738			36,738	
R^2 / Ω_0^2		.13 / .12			.09 / .08	

[3]While estimates of lower-order effects of race and utterance position are estimated using effects coding (black= -1, white= 1) in the body of the paper, we dummy code race here (black= 0, white= 1) for consistency with other models reported in this supplement.

"officer Respect increased more quickly in interactions with white drivers..."

How might this be restated as a causal question?

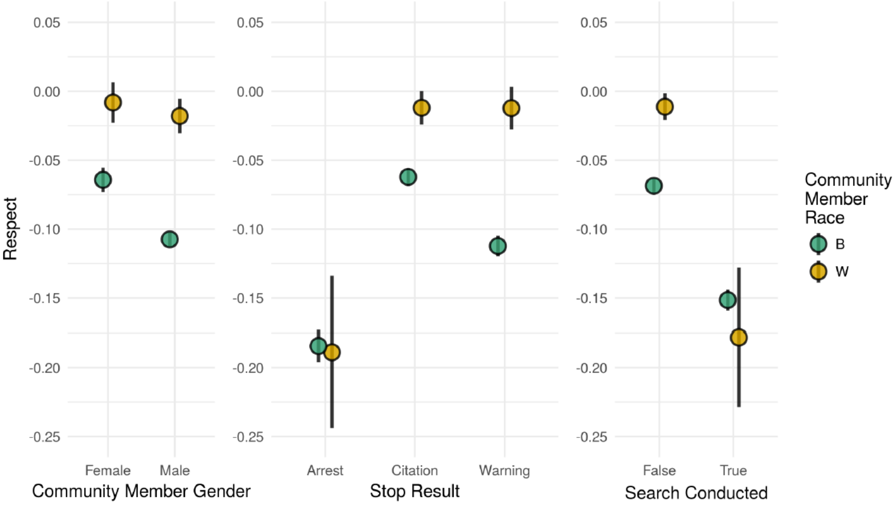
What is the effect of community member race on respect in officer language use?

$$\tau = \mathbf{E}(\text{Respect} \mid \text{do}(\text{Resident race} = \text{black})) - \mathbf{E}(\text{Respect} \mid \text{do}(\text{Resident race} = \text{white})) \quad (3)$$

Result of stop as post-treatment

$$
\begin{tikzcd}
\text{Actual offense} \arrow[dr] \arrow[drr, bend left] & & \\
\text{Resident race} \arrow[r] \arrow[rr, bend right] & \text{Officer respect} \arrow[r] & \text{Result of stop}
\end{tikzcd}
$$

But the descriptive raw means are compelling



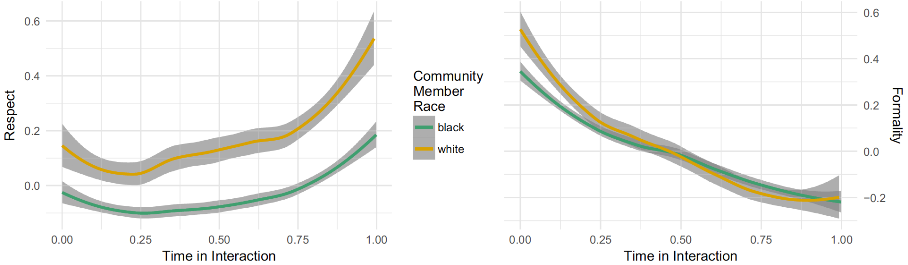


Fig. 5. Loess-smoothed estimates of the (*Left*) Respect and (*Right*) Formality of officers' utterances relative to the point in an interaction at which they occur. Respect tends to start low and increase over an interaction, whereas the opposite is true for Formality. The race discrepancy in Respect is consistent throughout the interactions in our dataset.

Linguistic classification accuracy of race

Mentioned briefly in first paragraph pg. 6525; pg. 13 of supplement

Similar logic to Gentzkow, Shapiro, and Taddy (2016)

- "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech."
- Use how easy it is to predict speaker party ID based on speech as a measure of political polarization
- The more predictive speech is, the greater polarization there is

In this paper

- Use how easy it is to predict the race of the community member being spoken to as a measure of racial disparity in officer language
- The more predictive officer speech is, the greater a disparity there is in how officers talk to black vs white residents

What they do in this paper

- Take a random balanced subsample of data (50% utterances directed at white residents, 50% directed at black residents)
- Extract same linguistic features as earlier + n-grams up to length 3
- Select 5000 most informative features
- Train a classifier using logistic regression to predict race based on these features

Mean predictive accuracy in 10-fold cross validation: **67.7%**