We some young kings:

Communities, age, and African American English online

Ian Stewart, Dartmouth College
ian.b.stewart.14@alum.dartmouth.org
Advised by Sravana Reddy, James Stanford

Motivation

The structure of AAE's syntax is well understood[1] but its demographic distribution remains debated. This study seeks to:

1. Develop a method to reliably extract short-range syntactic constructions from Twitter data via regular expressions.
2. Corroborate geographic[2] and/or demographic variation in syntax usage.

Data Collection

Geotagged Tweets (mined July - December 2013)

↓

Filtering (users > 20 tweets)

↓

228 million tweets, 1 million users

Content → POS Tags[3] → Regular Expressions

Metadata → ZIP Codes (linked to US Census demographic data)

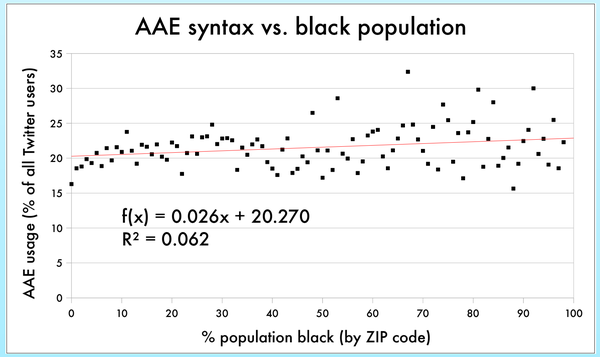
Metadata → Gender (estimated using 1995 Social Security Administration name database)

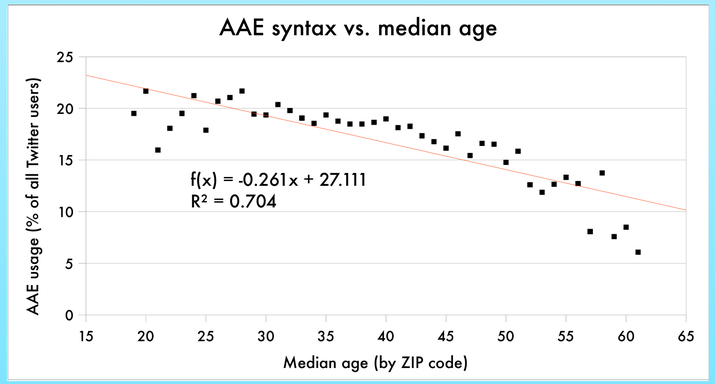
Construction[1]	Example	Frequency
ass camouflage construction (ACC)	*divorced her ass*	578,422
continuative *steady*	*steady washing her face*	15,979
copula deletion	*you mad*	1,545,024
future *finna*	*I'm not finna stress myself*	338,143
habitual *be*	*he be drinking that beer*	588,752
negative concord	*I ain't got no time*	455,531
negative inversion	*ain't nobody scared of you*	97,042
null genitive	*everybody & they mama*	296,358
past completive *done*	*Hunter done killed her*	87,346
remote past *been*	*Chris been told me*	16,998

Results

Ethnicity?

- Weak positive correlation between black population and AAE syntax usage.
- Unreliable trends for other ethnic demographics and geography (ex. by state).



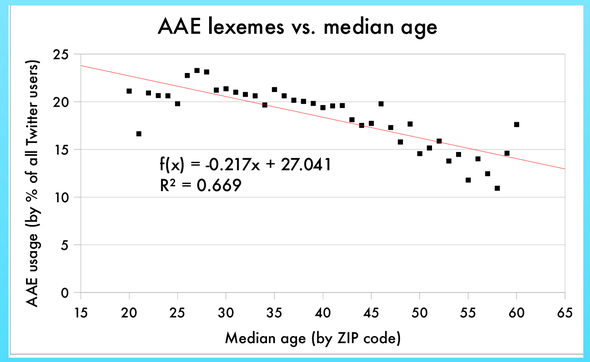


Median Age!

- Clear negative correlation, especially from ages 20-60 (>95% of data).
- Growth in usage through 20s followed by steady decline.

Confirmation.

- Same test performed for AAE lexemes[4] (ex. <talmbout>) and baseline lexemes (ex. <the>).
- Identical negative trend in AAE lexemes and absence in baseline lexemes.



Conclusions and Future Directions

- No significant variation in AAE syntax usage with respect to ethnic demographics.
  - Potential cause: African American users' overrepresentation in Twitter[5].
- Strongest trend? Median age.
  - Community-level curvilinear principle, even outside of towns with expected Twitter representation.
  - Age might be correlated with social structure (college town vs. family suburb).
- Caveat: hard to separate legitimate language use from quotations.
- Next step: check for connections between median age and other community patterns.
- Determine importance of social networks versus community statistics.
  - Twitter users often group by ethnicity[6].
  - Denser social networks in younger communities?
- Improve syntax detection accuracy.
  - 16.6% false positive rate, incalculable true/false negatives.
  - More rigorous regexes, possibly with crowdsourcing.
  - Training word-vector model for semantics (ex. ACC).
  - Bootstrap syntax from lexicon (ex. train on <talmbout> tweets).

References

1. Rickford, John R. *African American Vernacular English*. Malden, MA: Blackwell Publishers, 1999.
2. Wolfram, Walt. "Urban African American Vernacular English: morphology and syntax." *A handbook of varieties of English*. Ed. Bernard Kortmann. Walter de Gruyter, 2004.
3. Olutobi Owoputi et al. "Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters." *Proceedings of NAACL-HLT 2013*. Atlanta, GA: 2013.
4. Ofori-Atta, Akoto. "Here's Your Black Twitter Welcome Manual" *The Root*, Jan. 3 2014.
5. Duggan, Maeve and Joanna Brenner. "The Demographics of Social Media Users - 2012." Washington, D.C.: Pew Research Center, 2013.
6. Shane Bergsma et al. "Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter." *Proceedings of NAACL-HLT 2013*, Atlanta GA, 2013.