Investigating Morphosyntactic Variation in African American English on Twitter*

Tessa Masis *(they/them)*[1] **Chloe Eggleston**[1] **Lisa Green**[1]
Taylor Jones[2] **Meghan Armstrong**[1] **Brendan O'Connor**[1]
[1]University of Massachusetts Amherst, MA, USA
[2]Naval Postgraduate School, CA, USA
{tmasis,brenocon}@cs.umass.edu cegglesto@umass.edu
lgreen@linguist.umass.edu thelanguagejones@gmail.com
armstrong@spanport.umass.edu

African American English (AAE) is a language variety primarily spoken by most African American people in the United States and, like many languages, can vary regionally, stylistically, and generationally. However, early work on AAE perpetuated myths that the language variety was uniform across regions and that it was spoken primarily by working class men, due to being conducted in inner city areas and examining a specific set of linguistic features – such as the negative concord feature e.g. *I ain't done nothing like that before* (Wolfram, 2007; Wolfram and Kohn, 2015). These sociolinguistic myths negatively impacted not only the field of linguistics but also how the public viewed AAE (Wassink and Curzan, 2004). Since then studies have looked at a broader range of geographical areas and demonstrated distinct local differences. Here we build on this line of research by analyzing relative incidences of 18 morphosyntactic features (selected from Green (2002) and Koenecke et al. (2020)) in relationship to geographic and social factors, at scale.

Our data is a corpus of 224M geotagged tweets, posted across the entirety of the United States between May 2011 and April 2015 and filtered to prioritize conversational language. This dataset is five orders of magnitude larger than previous social media studies of AAE (Jones, 2015; Austen, 2017; Ilbury, 2020) with at least some data in all U.S. counties.

Many feature-based studies of large corpora use keyword searches or regular expressions to detect features – however, keyword searches are limited by orthographic variation in tweets and regular expressions cannot be made for all features. To circumvent these obstacles, we use the BERT-based machine learning method used in Masis et al. (2022) to automatically detect features. A binary classifier is trained for each morphosyntactic feature by fine-tuning a large pretrained language model; given a tweet, each classifier returns a score indicating the probability that the tweet contains the given feature. We use relative incidence - percentage of tweets containing the feature out of total tweets - to represent usage frequency. For each feature, relative incidence z-scores were calculated for all census tracts. Following this, Principal Components Analysis was used to identify common patterns of variation across the linguistic features (Grieve et al., 2011) and the first principal component (PC1) was shown to correspond to a latent factor of general AAE. We investigated the relationship between PC1 and 10 demographic variables (using data from the American Community Survey) via a standardized linear regression analysis, allowing us to explore the effects of demographic variables on general AAE usage while accounting for potentially confounding variables.

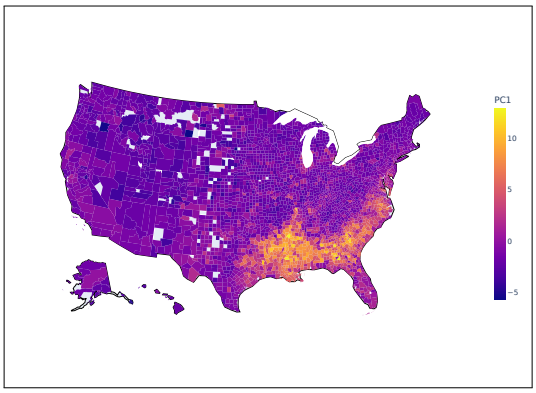

Figure 1: Heatmap of PC1, our latent factor of general AAE morphosyntax. Counties with sparse twitter data were excluded (in gray; ∼3%). County-level data is used here for visualization purposes; we use census tract-level data for the main analysis.

Our results show that, contrary to sociolinguistic myths of uniformity, there is clear variation in AAE

*This abstract contains material previously presented at NWAV50 in October 2022, and a version similar to the current will be presented at IC2S2'23 in July 2023. Our thanks to the reviewers of these conferences for their helpful comments.

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 392-393.
Amherst, Massachusetts, June 15-17, 2023

	Northeast		South		Midwest	
	Metro	Non-metro	Metro	Non-metro	Metro	Non-metro
Number of tracts	884	32	2612	494	876	60
Average PC1 score	0.432	0.456	1.335	**2.618**	1.143	0.809

Table 1: Table showing average PC1 scores for metro vs non-metro tracts (as defined by the Rural-Urban Commuting Area Codes) in the Northeast, South, and Midwest regions (as defined by the U.S. Census); we see a clear locus of AAE in the non-metro South. All tracts included in this table have a similar relative African American population (15-25%) in order to control for African American population as a potential confounding variable.

across both geographic and social dimensions. We present multiple notable findings. Regionally, we see a distinct spatially contiguous southern core (Fig. 1) which aligns with national-level phonological and lexical variation in AAE, although it is less variable (Austen, 2017; Jones, 2020). Across social groups, there is higher AAE usage in the rural south (Table 1) and in Black-Hispanic contact communities – both of which are groups currently underrepresented in the literature and completely unrepresented in early work on AAE. We confirm here that there is a great need for scholarly attention towards these communities, as our results demonstrate that they may be loci of AAE.

This work provides a significant advance in descriptive work on AAE morphosyntax, presenting the first national-level description and analysis of overall grammatical variation in AAE in order to answer key questions about variation in AAE. More broadly, our methods demonstrate how machine learning tools can be applied to large-scale real-world data to help us gain a more representative understanding of language in marginalized communities.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and feedback. This work was supported by National Science Foundation grants 1845576 and 2042939; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Martha Austen. 2017. "Put the Groceries Up": Comparing Black and White Regional Variation. *American Speech*, 92(3):298–320.

Lisa J Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.

Jack Grieve, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23(2):193–221.

Christian Ilbury. 2020. "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of sociolinguistics*, 24(2):245–264.

Taylor Jones. 2015. Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter". *American Speech*, 90(4):403–440.

Taylor Jones. 2020. *Variation in African American English: The great migration and regional differentiation*. Ph.D. thesis, University of Pennsylvania.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Tessa Masis, Anissa Neal, Lisa Green, and Brendan O'Connor. 2022. Corpus-Guided Contrast Sets for Morphosyntactic Feature Detection in Low-Resource English Varieties. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 11–25, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Alicia Beckford Wassink and Anne Curzan. 2004. Addressing ideologies around African American English.

Walt Wolfram. 2007. Sociolinguistic folklore in the study of African American English. *Language and Linguistics Compass*, 1(4):292–313.

Walt Wolfram and Mary E Kohn. 2015. Regionality in the development of African American English. *The Oxford handbook of African American language*, pages 140–160.