9th International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Investigating grammatical variation in African American English on Twitter

Keywords: machine learning, natural language processing, sociolinguistics, language variation, social media

Extended Abstract

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* [10, 11]. These sociolinguistic myths negatively impacted not only the field of linguistics but also how the public viewed AAE [9]. 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 grammatical features (selected from [2, 7]) 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 [5, 1, 4] 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 a BERT-based machine learning method to automatically detect features [8]. A binary classifier is trained for each grammatical 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 [3] 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.

Our results show that, contrary to sociolinguistic myths of uniformity, there is clear variation in AAE 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 [1, 6]. 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.

References

[1] AUSTEN, M. "Put the Groceries Up": Comparing Black and White Regional Variation. *American Speech 92*, 3 (08 2017), 298–320.

[2] GREEN, L. J. *African American English: A Linguistic Introduction*. Cambridge University Press, 2002.

[3] GRIEVE, J., SPEELMAN, D., AND GEERAERTS, D. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change 23*, 2 (2011), 193–221.

[4] ILBURY, C. "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of sociolinguistics 24*, 2 (2020), 245–264.

[5] JONES, T. Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter". *American Speech 90*, 4 (11 2015), 403–440.

[6] JONES, T. *Variation in African American English: The great migration and regional differentiation*. PhD thesis, University of Pennsylvania, 2020.

[7] KOENECKE, A., NAM, A., LAKE, E., NUDELL, J., QUARTEY, M., MENGESHA, Z., TOUPS, C., RICKFORD, J. R., JURAFSKY, D., AND GOEL, S. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences 117*, 14 (2020), 7684–7689.

[8] MASIS, T., NEAL, A., GREEN, L., AND O'CONNOR, B. 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* (Gyeongju, Republic of Korea, Oct. 2022), International Conference on Computational Linguistics, pp. 11–25.

[9] WASSINK, A. B., AND CURZAN, A. Addressing ideologies around African American English, 2004.

[10] WOLFRAM, W. Sociolinguistic folklore in the study of African American English. *Language and Linguistics Compass 1*, 4 (2007), 292–313.

[11] WOLFRAM, W., AND KOHN, M. E. Regionality in the development of African American English. *The Oxford handbook of African American language* (2015), 140–160.

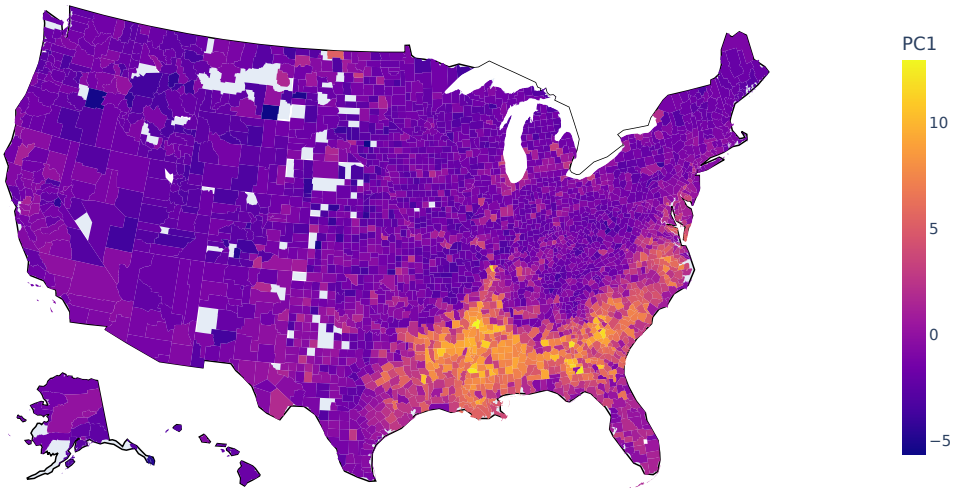

Figure 1: Heatmap of PC1, our latent factor of general AAE. Counties with sparse twitter data were excluded (in gray; ∼3%). County-level data is used here for visualization purposes; we use census tract-level data for the main analysis.

	Northeast		South		Midwest	
	Metro	Non-metro	Metro	Non-metro	Metro	Non-metro
Number of tracts	884	32	2612	494	876	60
Average PC1 score	0.4319	0.4595	1.3350	**2.6181**	1.1427	0.8089

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.