A large-scale Twitter-based exploration of morphosyntactic geographic variation in African American English

Early work on African American English (AAE) examined a specific set of features and was conducted in inner city areas, perpetuating the myths that there were certain features used by all speakers and that AAE was spoken only by working class people (Wolfram, 2007; Wolfram & Kohn, 2015). Since then studies have looked at a broader range of geographical areas, demonstrating distinct regional differences. We contribute to this line of research by using a corpus of Twitter data to analyze regional relative incidences for 21 morphosyntactic features, selected from Green (2002) and Koenecke et al. (2020), and present fine-grained national-level visualizations. Future steps will analyze feature co-occurrences; identify distinct geographic boundaries that reflect regional variation; and quantify correlation between feature usage and metrics that imply significant social factors – such as median household income, population density, or segregation index score.

Our data are 224M geotagged tweets from Twitter Decahose filtered to prioritize conversational language and limit automated posts, posted from the United States during May 2011 to April 2015. This dataset is five orders of magnitude larger than previous Twitter corpus studies of AAE, 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 novel BERT-based machine learning method to detect features, which extends a method used to detect Indian English features (Demszky et al., 2021) and has been used to replicate a number of manual linguistic analyses on CORAAL (Anonymous, under peer-review at COLING). 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 (software will be released after publication). We use relative incidence – percentage of tweets containing the feature out of total tweets – to represent usage frequency. Additionally, to take the ethnicity of Twitter users into account, we utilized data from the American Community Survey to weight each tweet according to the proportion of African Americans living in the blockgroup where the tweet was geotagged (Figure 1), which greatly varies by state, county, and even neighborhood. This weighting estimates feature usage conditioned on the aggregate proportion of African American speakers, without attempting to infer the race of individuals.

Based on a preliminary manual analysis of maps presenting relative incidences (Figure 2), we observe that there is a distinct Southern core, including rural areas (2a, 2b), which connects to the urban Midwest and Northeast via routes taken during the Great Migration (2c, 2d). This morphosyntactic regional variation aligns with national-level phonological and lexical variation in AAE with respect to frequency (Jones, 2015; Austen, 2017; Jones, 2020). Future work will shed light on the regional distribution of AAE features, the extent to which clusters of features occur in certain geographical areas, and which demographics use the features.

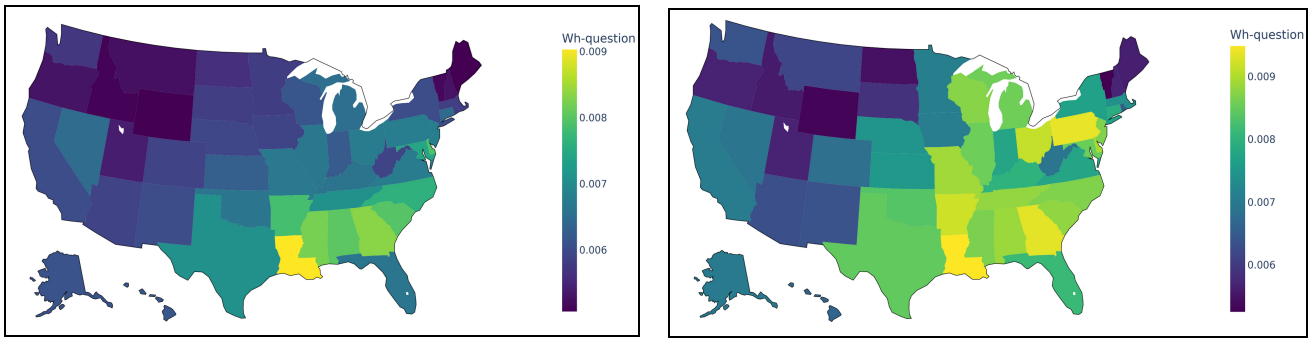
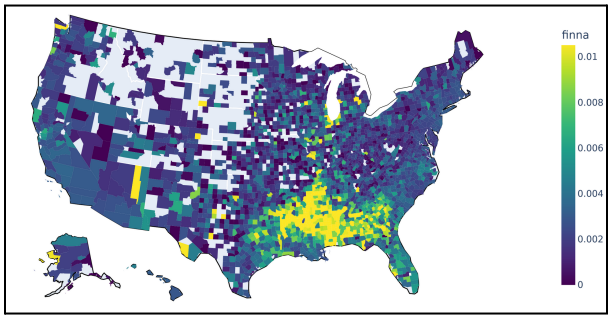
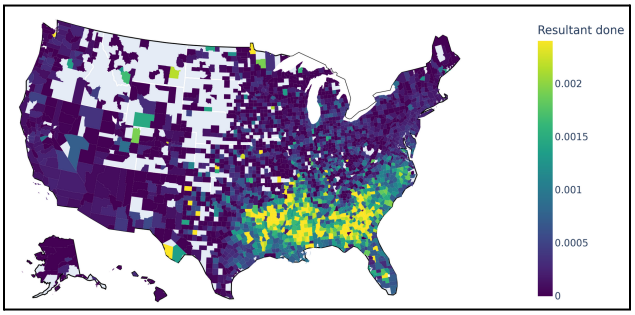


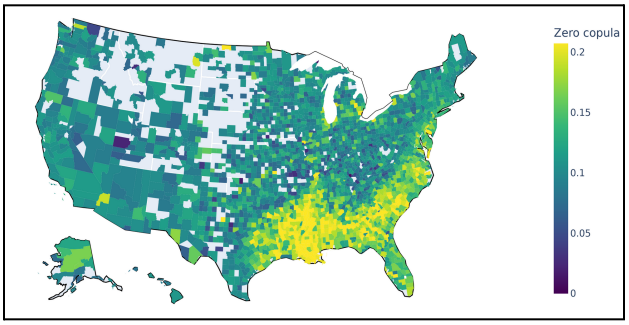
Figure 1: Relative incidence of *Wh*-question without demographic weighting (left), and with demographic weighting (right). Demographic weighting shows African American-conditional prevalence not just in the South, but also in the Northeast and Midwest.



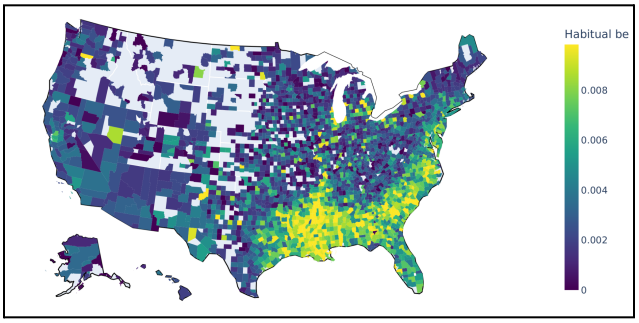
(a) Relative incidence of *finna*



(b) Relative incidence of resultant *done*



(c) Relative incidence of zero copula



(d) Relative incidence of habitual *be*

Figure 2: Relative per-county incidences of four features, all with demographic weighting. Counties with sparse data, due to low total tweets or low relative African American population, were excluded (in gray; ~12% of counties). Relative incidence of zero copula was calculated as the percentage of tweets containing zero copula out of total opportunities for zero copula to occur, in order to integrate the envelope of variation.

References

Anonymous. 2022. Corpus-Guided Contrast Sets for Morphosyntactic Feature Detection in Low-Resource English Varieties. Under peer review at COLING.

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

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to Recognize Dialect Features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

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

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. *Publicly Accessible Penn Dissertations*. 3929.

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.

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*, 140-160.