Chapter 2

Ain't + infinitive verb in Black/African American English

Kimberley Baxter[a] & Jonathan Stevenson[b]

[a]New York University [b]University of York

This study documents the first stage in the creation of an atlas of African American English (AAE) syntax, charting the relative use of *ain't* against *didn't* where they occur with infinitival verbs (*ain't see*/*ask*/*buy* etc. vs *didn't see*/*ask*/*buy* etc.). The source data is a large, geo-tagged Twitter corpus spanning three years from 2012–2015. The data are plotted for the contiguous United States, but the focus of this paper is on California, Illinois and Georgia. Previous literature (Fisher 2022, Kautzsch 2012, Weldon 1994, Howe 2005, Wolfram & Thomas 2002) notes that the use of *ain't*+infinitive (*ain't*+inf) structures is strongly linked to urban AAE speech communities and that the increased use of *ain't*+inf in rural AAE speech communities is likely due to interaction with urban AAE speech communities. The present paper re-examines the link between *ain't*+inf and AAE speech communities as they appear on social media via the use of Twitter data in a corpus-based analysis of AAE.

We present a method through which *ain't*+inf structures are isolated from other uses of *ain't* (e.g. *ain't-for-isn't*, *ain't+perfective*, etc.) and subsequently compared to *didn't+infinitive* (*didn't*+inf) structures in the same grammatical settings. These results are then compared to demographic information from the US Census and linked to the geographical metadata contained within each tweet.

We find that the strong link between *ain't*+inf and AAE speech communities is mirrored in language use on Twitter for Illinois, confirming results garnered using traditional methods. We find that *ain't*+inf is now prevalent in Georgia, indicating a near complete spread of *ain't*+inf constructions from urban centers in northern areas of the United States to southern areas. Meanwhile, we find that there is little to no correlation between *ain't*+inf use and high-density populations of Black/African American people in California, suggesting that *ain't*+inf has not spread to western coastal areas.

Kimberley Baxter & Jonathan Stevenson. 2025. *Ain't* + infinitive verb in Black/African American English. In Susanne Wagner & Ulrike Stange-Hundsdörfer (eds.), *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII*, 31–55. Berlin: Language Science Press. DOI: 10.5281/zenodo.15006595

1 Introduction

African American English (AAE) is one of the most widely studied varieties of American English. To date, there are numerous sociolinguistic studies that show how AAE syntax varies regionally (Weldon 1994, Moody 2011, Kautzsch 2012, Fisher 2022), thereby challenging earlier myths of AAE as a linguistic monolith (Wolfram 2007). The current paper adds to the growing body of research on syntactic variation in AAE by examining the relative use of *ain't* in conjunction with the infinitival form of a verb (*ain't*+inf) to mainstream *didn't* (*didn't*+inf) (1) on Twitter, across geographical space, focusing on California, Illinois and Georgia.[1]

(1) *ain't*+inf
   a. I ain't see her yet
      'I didn't see her yet.'
   b. She ain't say that
      'She didn't say that.'

The present paper re-examines the link between *ain't*+inf and Black/African-American (BAA) communities as defined by the US Census via the use of geo-tagged Twitter data in a corpus-based analysis of AAE.[2] Widespread use of AAE on archived social media posts creates a living database of timed, dated, and geo-tagged utterances from which corpora may be built.

2 On the use of Twitter data

2.1 Twitter demographics

As mentioned in Baxter (2025 [this volume]), tracking variation in a sociolect such as AAE is difficult on social media sites which grant anonymity to users. For Twitter, with the exception of verified profiles (indicated with a blue check), confirmation of one's race, ethnicity, gender, or other facets of one's identity are

[1] *Ain't*+inf is referred to as *ain't-for-didn't* in (Fisher 2022).

[2] The US Census does not account for differences in ethnicity among Black people in the United States – instead, all Black-identifying people are encompassed within the same singular category, *Black/African-American*. This category only seeks specification regarding whether one is "Black/African-American alone" or "in combination" with other races/ethnicities. This is an issue which affects all studies of AAE which seek to use the US Census as a touchstone for demographic metadata, including the present study. We use *Black/African American* (BAA) when referring to racial and ethnic identity in this chapter.

optional. As a result, a reliable mass verification of users' demographic data is not currently feasible.[3]

Multiple surveys have been conducted in an effort to tease apart the demographics of Twitter users, with varying results. The present study refers to the Pew Research Study "Sizing up Twitter" (Mitchell 2019). This survey is a representative sample of 2791 adult Twitter users surveyed via Ipsos KnowledgePanel, a probability-based online panel of US adults.

The survey found that, at the time of the study, approximately 80% of all tweets were made by about 10% of users. Twitter users tended to be younger, and more likely to be Democrat or otherwise left-leaning, more likely to be women, and more likely to tweet about politics. 11% of Twitter users identified as Black* (not including Hispanics). Of these, approximately 30% were college graduates, 40% had some college education, and 30% had a high school diploma or had not completed their high school education.

The survey did not include statistics on political affiliation, age, or use by race. However, the study does present a representative Twitter demographic which shows a proportion of Black tweeters which is comparable to the general population of the United States (approximately 13% of the total population of the United States, US Census Bureau 2018). The US Census Bureau recorded 22.5% of the Black population in the United States 25 years and older as having a bachelor's degree in 2015 (Ryan & Bauman 2016). This number is lower than the number in "Sizing Up Twitter" and appears to show that the finding that Twitter users are more likely to be college educated holds among Black users as an isolated demographic.

2.2 Twitter location data

A concern frequently raised with the use of Twitter data for linguistic research is that a tweet's geolocation (discussed in more detail in Section 4.3) reflects where the tweeter was located when they sent the tweet rather than where the tweeter is necessarily from.

However, in this research, we do not assume that any individual user is necessarily a long-standing resident of the place from which they tweeted, rather we take the aggregate of a large number of tweets from different users as indicative of the overall picture for that location. This works on the prediction that, for the most part, Twitter users are, in fact, tweeting from the location in which they reside.

[3]There have been attempts to estimate user age and gender using crowd-sourcing, see Nguyen et al. (2014), with mixed results.

This prediction is borne out in a number of studies which show that Twitter data tallies closely with data drawn through traditional methods. For example, Stevenson (2016) shows that geolocated tweets for ditransitive verbs with pronominal objects in the UK (*send it me*/*send me it* etc.) correspond closely to data from the Survey of English Dialects (Orton & Dieth 1962), and Strelluf (2019) shows a similarly close correspondence for positive *anymore* between Twitter data and established distributions.

2.3 Summary

With these issues in mind, we would nevertheless expect that if there is an association between *ain't*+inf and communities with high-density BAA populations (Fisher 2022, Labov et al. 1968, Kautzsch 2012, Weldon 1994) then this will also be reflected in Twitter data. Conversely, if there is no such association in the Twitter data for a given location, this would be evidence that either:

I. AAE language patterns documented in previous research via audio recordings, interviews, and other traditional data collection methods are not readily visible in data taken from Twitter.

   or,

II. The correlation between *ain't*+inf and high-density BAA communities has diminished.

However, as we will see, it is difficult to explain the systematic distribution of the Twitter data if they are not tallied to actual language use. If there was a fundamental problem with Twitter data, then we would expect to see it across the board, and it would not explain the consistent geographical differences in Twitter use that we find between census tracts and between states.

As will be discussed in the next section, the ethnic identity of a given user is not our first question. Rather, the geographic distribution of a linguistic structure previously associated with the linguistic variety AAE, is the first question. The second question is whether this distribution correlates with the geographic distribution of BAA people in different locations.

The goal here is to both supplement previous research and offer a tool for future research. The hope is that the resulting atlas may be used to find areas that warrant further investigation via traditional, on-the-ground methods.

This, we believe, is a significant advancement. Rather than investigating a given location that may arise through convenience, happenstance or personal

connections, the promise here is for an overview that may reveal previously unknown patterns and locations of particular interest. An analogy might be to that of using arial photography to scan for areas of archeological interest, rather than relying on serendipitous finds.

3 Background

Due the relative anonymity of Twitter profiles, which makes it difficult to verify the ethnicities of the users who own them, this study follows a *linguistic grouping* approach (Horvath & Sankoff 1987), which first groups linguistic features associated with a given variety, then looks for correlations with external, social factors. Accordingly, linguistic features of AAE are collected and coded *before* sociological factors such as race and ethnicity, rather than following the traditional *sociological grouping* approach used in most traditional sociolinguistic studies.

> The terms social and linguistic grouping do not mean that sociological consideration predominate in one approach and linguistic concerns in the other, but only refer to the temporal order in which they enter into the statistical analysis. (Horvath & Sankoff 1987: 180)

The present study starts by choosing a linguistic variable, in this case *ain't*+inf, and calculating indices of use against mainstream *didn't*+inf, mapped across the contiguous United States via the geolocation metadata attached to each tweet. These indices are then compared to indices of BAA population in US Census tracts across the same geographical area. By doing so, we aim to provide a broad yet comprehensive view of usage rates in high-density populations of BAA people relative to low-density populations.

While *ain't* itself is common across many varieties of English, it is often considered that *ain't*+inf is a distinctive feature in AAE, where it occurs more frequently than in other varieties of English (Fisher 2022, Labov & Harris 1986, Kautzsch 2012). So, for the present study, following the Language-First model, this means *ain't*+inf tokens, established as being part of AAE, are extracted from the Twitter API prior to testing association with demographic data regarding ethnicity.

Ain't+inf is thought to have been recently innovated in northern urban centers and spread as a result of language contact with AAE speakers who moved there during the Great Migration (Fisher 2022). For example, previous research (Jørgensen et al. 2015) suggests that there is an increased use of *ain't*+inf both diachronically and over apparent time by comparing early AAE recordings to later

AAE recordings, or by comparing the speech of Northern (Philadelphia) speakers to speakers who had moved to the Northern United States from the Southern United States.

The scale of the data available via Twitter's API allows the relative frequency of *ain't*+inf to be mapped across the United States and, at the same time, to provide unprecedented resolution at the level of small towns and suburbs. Furthermore, in-line with previous studies that use Twitter data for dialect research (Jones 2015a, Stevenson 2016, Willis 2020, Strelluf 2019, 2020), results in many cases appear to follow established dialect "faultlines" (Eisenstein 2013: 1) while also highlighting particular hotspots of use. If *ain't*+inf is unique to AAE, then we would predict that its distribution would correlate with BAA population distribution.

Principally, then, our main question is:

> To what extent is the association between *ain't*+inf and the BAA population reflected in data taken from Twitter?

While the broader study covers the entire US, the present chapter focuses most closely on *ain't*+inf use in Illinois, Georgia and California. These states were chosen because they represent three different regions of the United States, with California being on the west coast, Illinois being in the midwest, and Georgia on the southeastern coast, with all three housing cities which land among the top ten cities most densely populated by BAA people (Tamir et al. 2021). In addition, both Illinois and California were destinations to which many BAAs migrated during the Great Migration, and still have a high population density of BAA residents.

Finally, while our focus is on *ain't*+inf, we acknowledge that the next step in evaluating the extent to which this form of *ain't* is unique to AAE within geotagged Twitter data is to see how it patterns relative to other forms that are well attested in other Englishes. We discuss these next steps in Section 8.

4 Method

The semantic near-equivalence of the sentences in (1) allows us to consider the two forms as variants of a single variable (Labov et al. 1968, Wolfram & Schilling-Estes 2016, Fisher 2022) whereby the relative frequency of *ain't*+inf may be measured against *didn't*+inf to provide an index of use across and between places, without needing to know the overall corpus size for a given place.

In this way, *didn't*+inf provides a yardstick against which to measure *ain't*+inf usage via the *straight* model seen in Rickford et al. (1991), through which the use

of a given part of speech is divided by all potential outputs within that grammatical space. Where Rickford et al. (1991) presents the *straight* model with reference to the calculation of copula usage and deletion, we present a similar *straight* model for *ain't*+inf which divides the frequency of *ain't*+inf occurrences by the sum of *ain't*+inf and *didn't*+inf. The resulting number is hereafter referred to as the *ain't+inf index*:

$$ain't{+}inf\ index = \frac{ain't{+}inf}{(ain't{+}inf + didn't{+}inf)}$$

The next step, then, is to extract instances of *ain't*+inf and *didn't*+inf from the Twitter API.

4.1 Extracting data using the Academic Twitter API

The Academic Twitter API (ACTW) was used to extract Twitter data. ACTW was made available at the start of 2021 and permits academic access to the entire Twitter archive at no cost.

To extract tweets containing *ain't*+inf and *didn't*+inf, a simple script was written in R to generate a list of strings combining an infinitival verb, from a list of 150 common verbs, with either *ain't* or *didn't* – both with, or without the apostrophe (*aint*, *ain't*, *didnt* and *didn't*).[4] The result was a list of 600 strings (4 × 150):

```
aint see
ain't see
didnt see
didn't see
aint ask
ain't ask
didn't ask
didnt ask
```

The resulting list was then used to generate a search query formatted for the Twitter API that could be used to pull tweets for each item in the list. In the example below, *q* represents a given string combination for that stage of

[4]While it would be search for all possible infinitive verbs occurring with *ain't* or *didn't*, using the top 150 most common verbs would capture the vast majority of cases. It is assumed that the inclusion of less frequently occurring verbs would not significantly sway the results, in aggregate.

a *for loop*.[5] The search query also disallowed retweets (-is:retweet), and only tweets that contained geolocation metadata (has.geo). In sum, Twitter is searched for each string in turn until all strings have been searched. Using this method, approximately 3.2 million geolocated tweets were collected, each containing a variation of either *ain't*+inf or *didn't*+inf.

```
query <- paste0('(',q,')', ' has:geo -is:retweet place_country:US'
```

4.2 Data cleaning

The first step in cleaning the data was to remove results where punctuation intervenes between *ain't* and the verb. Twitter's search engine is blind to punctuation, so the raw data will include strings such as in (2).

(2) But he aint. Ask someone else

Removing these was done using a simple regular expression search in R, which is sensitive to punctuation. A table was generated containing all instances of the strings in the initial search coupled with additional coding data such as verb type and whether it was *ain't* or *didn't*. When the Twitter data was matched to an entry in the table, the additional coding data could also be carried over, resulting in a coded dataset.

Next, the data were spot-checked for further false positives. Problematic cases where the infinitive form is homonymous with another part of speech, which can readily occur in same position in the sentence, were investigated. In some cases – such as: *like* as in *she ain't like me*; *love* as in *that aint love* and *fly* as in *she aint fly* – the non-infinitival form was so dominant and removed entirely. This was only necessary for a handful of verbs, however.

4.3 Location data

Each tweet comes packaged together with various forms of associated metadata. Of this metadata, there are three main types directly linked to the user and the tweet location.[6]

[5] A *for loop* is a function used in computer programming that performs a set of sub-functions a given number of times. In this case, the *for loop* performs a Twitter search for every construction in the given list of constructions.

[6] There are other ways to garner user and tweet location from the content of a user's tweets, or other information that they provide. For example, a user may state in a previous message that they grew up in Chicago or went to school in South Atlanta. This method of gathering additional location data is particularly useful when data is more scarce, such as when studying

First, the GPS points for the actual location that the tweet was sent from. Evidently, this data is only available for tweets sent from mobile phones. Additionally, GPS data is only available for a relatively small fraction of Twitter messages, around 1–3% by most estimates and data is most prevalent prior to 2015 when a setting in the phone application – to add GPS data to each tweet – was opt-out rather than opt-in. GPS data comes as a set of coordinates:

```
c(-76.0792, 38.566)
```

Second, there is user entered location data that is free form, and associated with the user profile itself. This data is information provided by the user when they set up their account, but can be changed later. It is not required and can be any kind of free-entered text. This data can be useful, but may be unreliable and may not correspond to a place at all – it may be used to indicate personality type, for example.

```
"Trill, Texas"
"VonteWorld///Beast Coast"
"Moss Bluff, LA"
"at the bar"
"INFP"
```

Third, tweets also contain a place.id, which is a code corresponding to a rectangle defined by four geographic coordinates, corresponding to a place as can be seen in Figure 1 (page 40).

The place to which the ID is assigned is derived by Twitter through a process of *data enrichment* using a combination of fuzzy matching of user entered location and GPS coordinates.

4.4 Twitter atlas, first iteration

It is possible to generate a useful atlas of tweets using the place.id alone, and indeed, the first iteration of the current atlas used this data. Here, the counts for each variant (*ain't*+inf and *didn't*+inf for each place.id were represented as pie charts on an interactive atlas. Pie charts were placed at the center point of the

a lesser used language like Welsh (Willis 2020) or to drill more data from a dataset (Gopal et al. 2021). In addition, taking the location of where a user grew up is closer to the methodology employed in traditional dialectology, and may, in some sense, more authentically reflect the language of a given place, though see Stevenson (forthcoming). This method is, however, less straightforward for the current investigation, in the US, where many places share the name, though there are ways to mitigate this issue. For now, we leave this to future study.

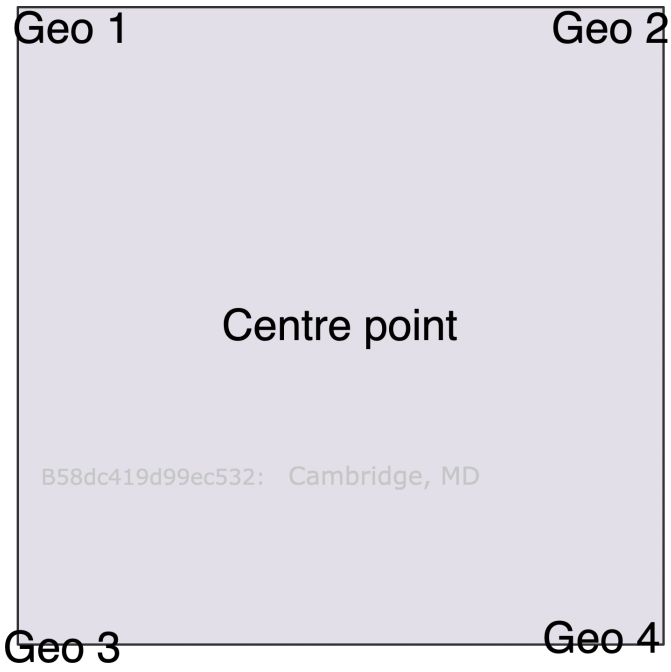

Figure 1: Illustration of Twitter place.id, a square defined by four geocode points, corresponding to a given place.

place.id area (see Figure 1). This was done using the *leafletminicharts* package for R (Bachelier et al. 2021). Using this package it was also possible to add a level of interactivity where the user can click on a pie chart and see a breakdown of the exact counts for that location. The user can then click on a variable and see a sample of the actual data from that place.

This was a valuable tool at this stage in the research process, allowing us to investigate different places, and offering a different way to probe the dataset by place. Figure 2 shows a screenshot of the interactive atlas focussing on the East North Central region of the US, specifically on the Chicago area.

In order to be represented on the map, it was set that a place.id would have to have at least 100 tweets associated with it. This was done to ensure that enough data was present to draw statistically significant conclusions. In addition, this reduced the number of pie charts that had to be rendered on the map, improving legibility and performance (setting to lower that 100 per place meant that scrolling the atlas became too slow).[7]

[7]We should note here that the same limited dataset was used for the second iteration of the atlas, which was not necessary for that stage. We do not believe that this will affect the results for the second stage but future versions will have this legacy limit removed.

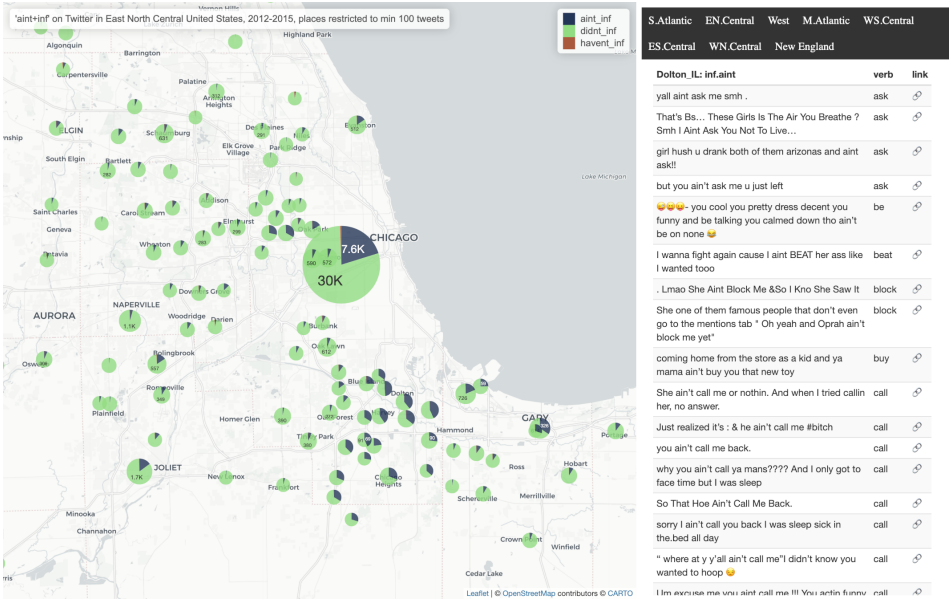

Figure 2: Screenshot of the first iteration of the interactive atlas that uses Twitter place.ids rather than GPS points. Focus here is on the Chicago area. Interactive version available at http://nwdialectatlas.uk/infaint/.

It became apparent immediately that there was geographic variation in the use of *ain't*+inf versus *didn't*+inf. Closer examination of places with high *ain't*+inf use showed a pattern: all had high BAA populations. In Figure 2, it is possible to see southern Chicago and Gary with high rates of *ain't*+inf, represented in blue, both areas with high BAA populations, meanwhile northern and western areas have very low rates of *ain't*+inf, which follows if *ain't*+inf is distinctive to AAE.

4.5 Twitter atlas, second iteration: linking to census data

In order to show a systematic link between *ain't*+inf use and BAA population, it is necessary to link Twitter place.id to demographic data. In the US, the most reliable demographic data is provided as part of the census. Census data is organised by tracts, with each defined by geographical polygons, stored as shapefiles. Each tract is associated with demographic information, such as race, about that geographic location. Tracts are defined for the purpose of administering the census, sometimes following political boundaries. They are the smallest geographic areas that are recorded with associated metadata.

However, it is not possible to directly link the Twitter place.id areas to tracts. Sometimes a place.id will encompass numerous tracts and other times, a census tract will contain several place.ids.

So, for the purposes of the present investigation, where the aim is to measure *ain't*+inf against demographic data, it is necessary to have Twitter location data that is compatible with the location information provided for the US census. Associating tweets with census tracts was done using the point-in-polygon technique (similar to Blodgett et al. 2016). This technique, illustrated in Figure 3 is a method for calculating whether a given point falls within the bounds of a polygonal shape. In Figure 3, the red points have been found to be within the polygon.

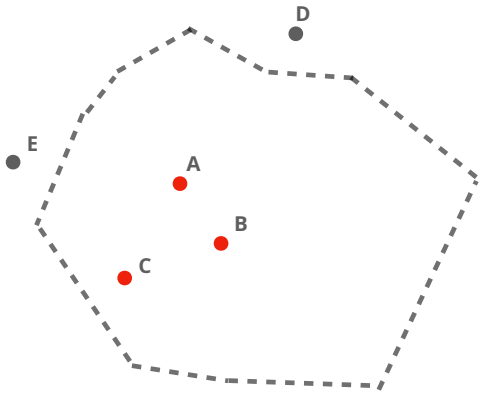

Figure 3: Illustration of point-in-polygon technique. Source: https://datawanderings.com/2018/09/01/r-point-in-polygon-a-mathematical-cookie-cutter/

The PinP technique was used to plot from which census tract a given tweet was sent. As we already know the racial demographic data for each tract, then the relative frequency of *didn't*+inf and *ain't*+inf (*ain't+inf index*) could be measured against the relative population of BAA against the total population (*BAA index*).

4.6 Retrieving census data

US census data is freely available from census.gov and accessible directly in R using the *tidycensus* package (Walker & Herman 2023). Tidycensus makes it straightforward to retrieve US Census data that is prepared for use with other R packages in the *tidyverse* suite (Wickham et al. 2019). It also makes it relatively easy to work with shapefiles using the sf package. A shapefile is a standard format for storing geographic data as vectors, in terms of points, lines, and polygons.

For our purposes, it is possible to use tract data to calculate an index of the BAA population by dividing BAA population by the total population for each tract.

The index, hereafter referred to as the *BAA index*, is calculated by dividing the BAA population by the total population.

$$\text{BAA index} = \frac{\text{BAA population}}{\text{total population}}$$

This can be seen in the map presented in Figure 4.[8]

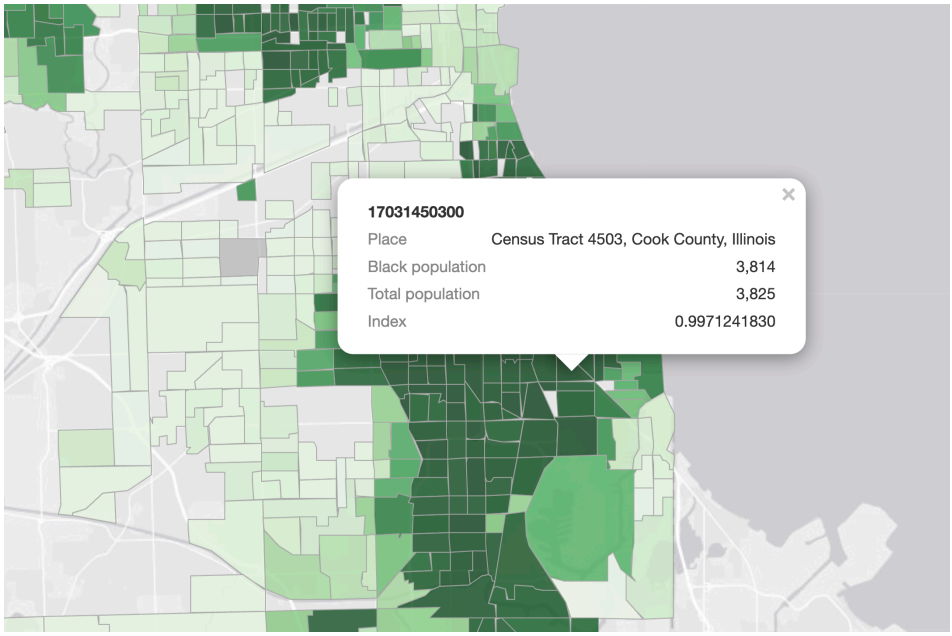

Figure 4: Cook County Census Tracts Tracts, based on data from US Census *American Community Survey* 2011–2015 (US Census Bureau 2015)

Figure 4 is a map of Chicago, Illinois. The maps are divided into census tracts, each of which are linked to demographic information taken from the United States Census. The text-box in the center of the map displays information for one of the census tracts within the city: Cook County. This text-box appears as a pop-up when a census tract is clicked by the user.

In the case of Cook County, nearly 99.7% of residents within this census tract are identified as BAA, which is reflected in the darker shade of green in this

[8]An interactive version of the maps presented in this chapter is available at: http://nwdialectatlas.uk/infaint/.

and surrounding census tracts. Tracts with low populations of BAA people are represented by lighter shades of green.

4.7 Mapping and plotting Twitter data with census data

First, using the *ain't+inf index* and *BAA index* for each tract, two maps of the US were produced – an *ain't*-usage-census-tract-map, and a racial-demographic-census-tract-map – which could be used for a visual side-by-side comparison. If *ain't*+inf is associated with BAA population groups, then we expect to see a visual similarity in the resulting maps.

Second, scatterplots were produced for each State, plotting *ain't+inf index* against *BAA index* with each point representing a single census tract. If *ain't*+inf is associated with BAA populations, then we expect to see a clear linear correlation between the two indices.

Finally, the data were broken down into high *BAA index* (>90%) and low *BAA index* (<10%) corresponding to high and low *BAA index*. Boxplots could then be produced for each group, with each State represented by one box. Again, if *ain't*+inf is unique to AAE, then we expect to see tracts with a high *BAA index* coalescing (represented by short boxes) at a high rate of *ain't*+inf and, conversely low *BAA index* with low *ain't*+inf.

5 Results

Here we present data on the three states under investigation: Illinois, Georgia and California. For each state, we start with heatmaps for Chicago, Atlanta, and Los Angeles and then present scatterplots for each respective state.

We then present a comparison between *ain't*+inf use across 13 states between high BAA population (>90%) and low BAA population (<10%).

5.1 Illinois

The two first maps, presented in Figure 5, compare BAA distribution (left) and *ain't*+inf distribution (right).

As stated, the distribution is represented by the *BAA index*, which is calculated by dividing the BAA population of a given census tract by the total population of that census tract.

From a simple observation of the two maps, it is evident that the prediction that *ain't*+inf is associated with BAA population is borne out. The most concentrated

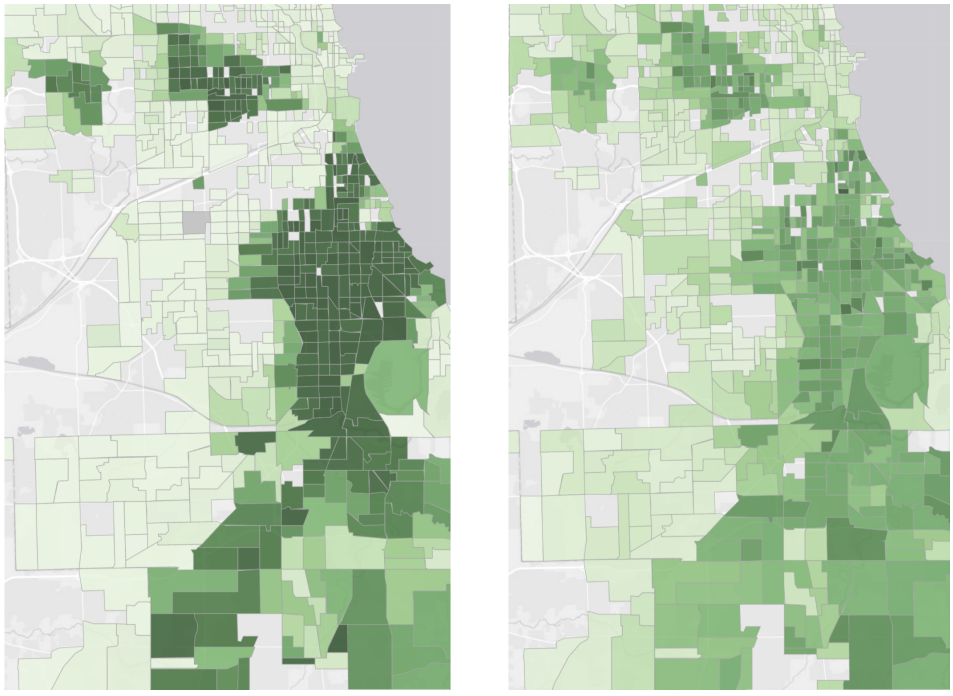

Figure 5: Chicago maps. Left = BAA Population Index, based on data from US Census *American Community Survey* 2011–2015 (US Census Bureau 2015); Right = *ain't+inf index*

areas of the *ain't*+inf map are also the most concentrated areas of the *BAA index* map.

This is further clarified in the scatterplot for Illinois, in Figure 6 which shows a clear correlation between *BAA index* and *ain't+inf index*; the higher the BAA population, the higher the *ain't+inf index*. The lower the BAA population, the lower the *ain't+inf index*.

Additionally, the segregation that exists in this part of the country is apparent, with large clusters of census tracts clustering at the top and bottom of the chart corresponding to tracts with nearly 100% of residents identifying as BAA, or nearly 0% BAA, while points in the center of the plot are more sparse, corresponding to there being far fewer tracts with a mixed population.

5.2 Georgia

The maps for Atlanta (Figure 7) again show a clear similarity in distribution, comparing *ain't*+inf and BAA populations.

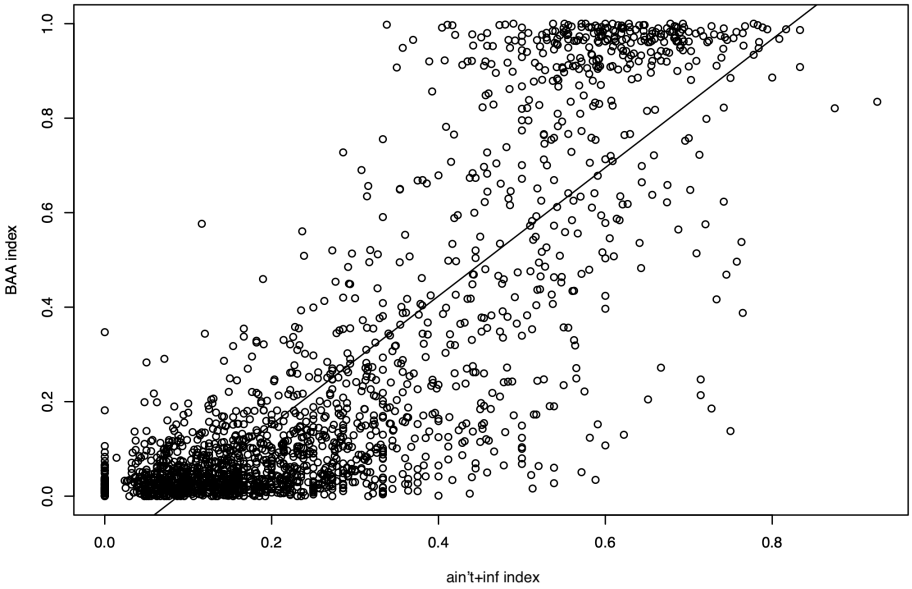

Figure 6: Scatter plot for Illinois: Vertical = BAA Population index; Horizontal = *ain't+inf index*

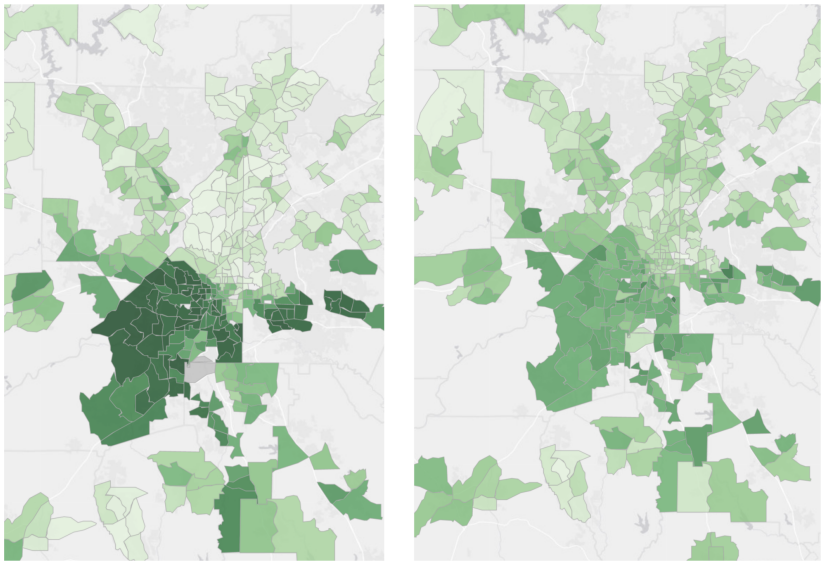

Figure 7: Atlanta maps. Left = BAA Population Index, based on data from US Census *American Community Survey* 2011–2015 (US Census Bureau 2015); Right = *ain't+inf index*

Meanwhile, the scatterplots for Georgia (Figure 8) contrast with those for Illinois (above) showing a smoother distribution of BAA population, while – crucially – still showing a clear correlation between *BAA index* and *ain't+inf index*.

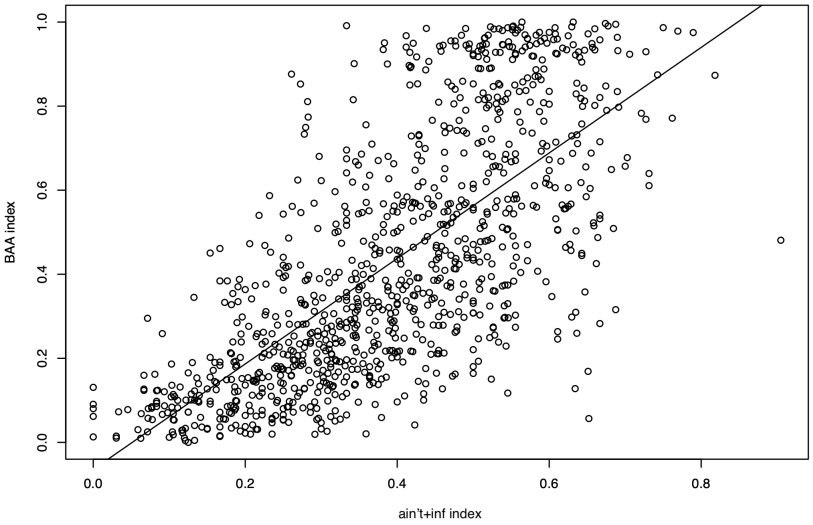

Figure 8: Scatter plot for Georgia: Vertical = BAA Population index; Horizontal = *ain't+inf index*

5.3 California

Particularly notable, however, is the absence of this correlation in the maps for Los Angeles, California (Figure 9). While Illinois and Georgia show a very clear correlation between census tracts with high BAA populations and high *ain't*+inf indices, results show that use of *ain't*+inf in California is not only markedly lower, but less concentrated.

The scatter plot in Figure 10 supports this weaker correlation between *BAA index* and the *ain't+inf index* in California. Where the plots of the census tracts for Georgia and Illinois both show clusters of census tracts with a *BAA index* of above 0.8, California shows a much lower number of census tracts which meet the same criteria.

Of the tracts that do show a *BAA index* of above 0.8, none show an *ain't+inf index* above 0.4. This is in contrast to Georgia and Illinois, where most tracts that show a *BAA index* of 0.8 or above also show an *ain't+inf index* of approximately 0.4 and above.

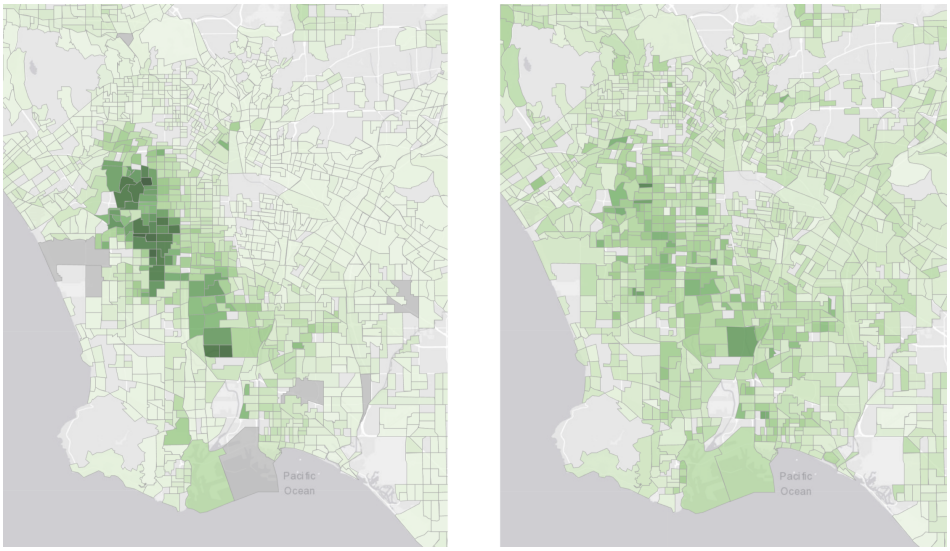

Figure 9: Los Angeles maps. Left = BAA Population Index, based on data from US Census *American Community Survey* 2011–2015 (US Census Bureau 2015); Right = *ain't+inf index*

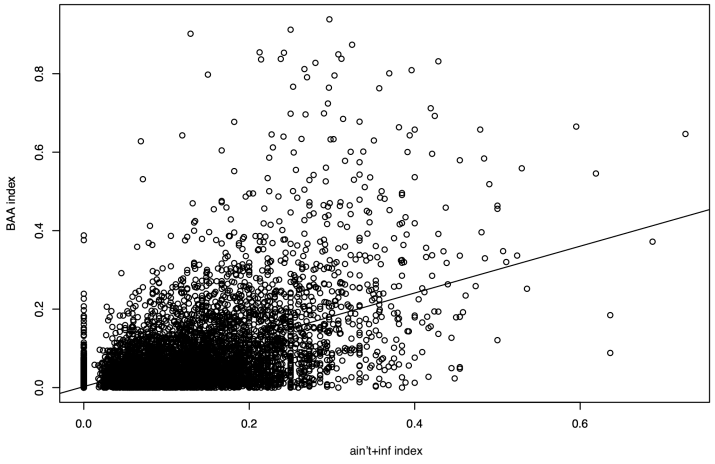

Figure 10: Scatter plot for California: Vertical = BAA Population index; Horizontal = *ain't+inf index*

5.4 Comparing states

Figures 11 and 12 are boxplots representing the *ain't*+inf indices across 13 states in census tracts with 10% or less BAA/BAA population and 90% or more BAA/BAA population, respectively, and shows that unlike other states, the *ain't+inf index* in California remains relatively low regardless of *BAA index*.

The boxplots in Figures 11 and 12 further underscore the marked difference in *ain't*+inf use in California, as compared to other states. Both in census tracts with 0.9 *BAA index* or greater and census tracts with 0.1 *BAA index* or less, the average *ain't+inf index* remains relatively low, and does not increase regardless of racial population density. Also notable are *ain't*+inf rates in Alabama (AL), Louisiana (LA), and South Carolina (SC), which all show slightly higher *ain't*+inf indices among census tracts with 0.1 *BAA index*, showing at around 25% usage where other states show >20% usage.

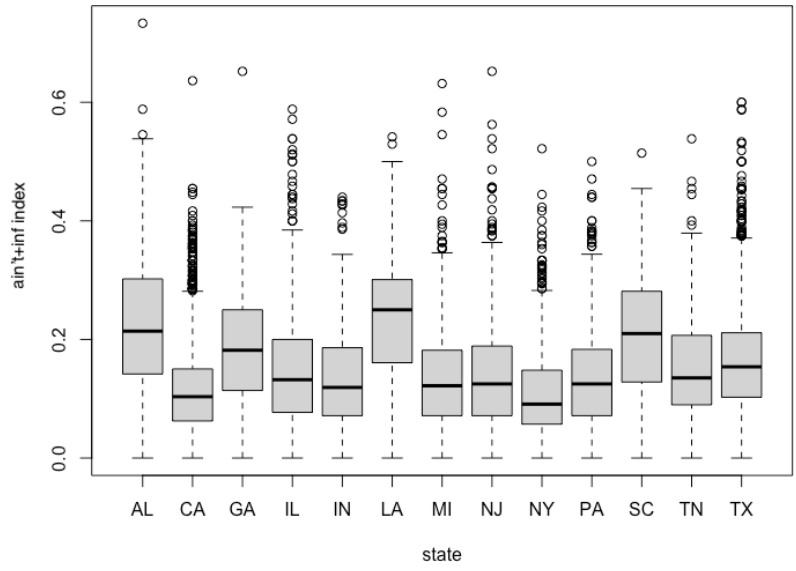

Figure 11: Boxplot showing *ain't+inf index* for tracts with <10% BAA population

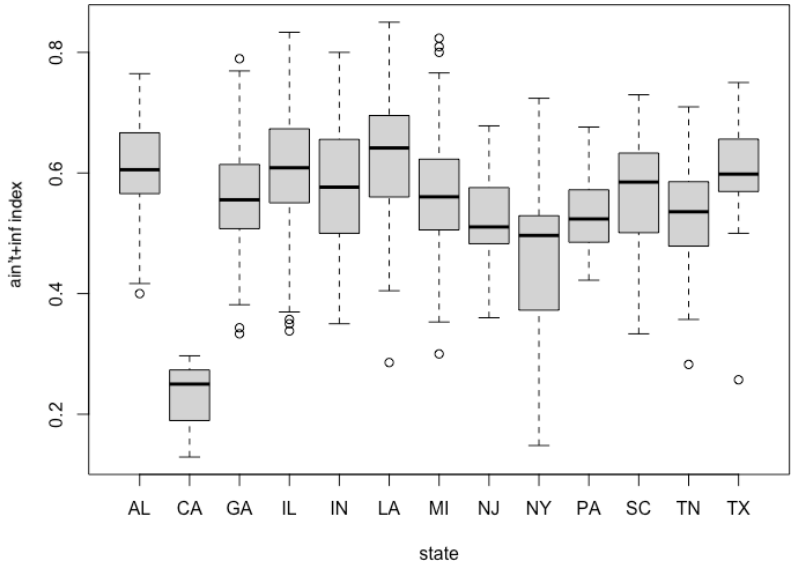

Figure 12: Boxplot showing *ain't+inf index* for tracts with >90% BAA population

6 Analysis

The results above confirm the conclusions reached in previous research, which found a correlation between the use of *ain't-for-didn't* constructions and BAA speakers. As a result, they also support previous findings (Eisenstein 2013, Jones 2015b, Willis 2020) that language use on social media platforms such as Twitter patterns with natural language distributions found using traditional methods.

Ain't+inf is considered to have been originally relatively rare in early AAE in the Southern US, with the structure having been innovated in the northern cities after the Great Migration. If this is indeed the case, the current distribution shows a continued and near-complete spread of *ain't*+inf within BAA communities in the United States from the northern urban centers to southern urban centers.

This is important to note when we consider Fisher (2022), whose analysis suggested that *ain't*+inf and other forms of *ain't-for-didn't* were used more frequently among speakers who were born and raised in Philadelphia than they were among speakers who had recently migrated from the South. The data also reveal that this spread has not reached as far as California, where *ain't*+inf is shown to be relatively infrequent, and uncorrelated with the distribution of BAA populations.

Why this North-South spread has occurred, and why it has not spread west to California is unclear at this stage, but is a topic ripe for future investigation. A first point of investigation here would be to analyse migration patterns from the northern cities to the south. Cities such as Atlanta have re-emerged in recent years as BAA cultural centers, attracting new waves of BAA migrations. It may also be partly explained by ongoing family and social ties between the north and south that have facilitated the spread of *ain't*+inf south, but not west. It will be interesting to see if *ain't*+inf is part-way through a process of diffusion, and will eventually reach California, or if it will remain a difference between the varieties of AAE in west and east.

7 Conclusion

This paper discussed variationist studies in African-American English and aimed to contribute to the study of syntactic variation therein by conducting a study of language data mined from Twitter via the Twitter API.

We presented an analysis which also used Twitter data to confirm analyses reached by more traditional methods in previous literature, thus showing the viability of Twitter data in the study of syntactic variation in AAE. To collect the data, we examined the parameters through which infinitival *ain't* occurred and isolated it from other uses of *ain't* which were more ambiguous.

To map the data, we calculated an index of infinitival *ain't* use and used a point-in-polygon approach to map the geospatial metadata within each tweet within census tracts taken from the US Census. The resulting maps show a correlation between infinitival *ain't* and census tracts with high BAA/African-American populations in Georgia and Illinois, but show a much weaker correlation, as well as a much lower average *ain't+inf index* in California. Data shows that this weaker correlation holds regardless of race.

The results of each map were confirmed by scatter plots and box plots which also showed a correlation between *BAA index* and *ain't+inf index* in communities with high populations of BAA/African-American identifying residents, according to the US Census, with the exception of California.

These results indicate a likely difference in both the frequency and manner of *ain't* usage in the United States, despite previous descriptions of the variety as uniform within urban centers. These results represent a starting point for future researchers to analyse them in more detail via alternate methods such as surveys and/or interviews.

Finally, we have shown how the language-first approach adopted here using Twitter data is compatible with, and reflective of, already-established conclusions gleaned from more traditional identity-first analyses.

8 Future directions

As stated at the outset, the work presented in this chapter represents the first step in what is planned to be a much larger atlas of AAE use using both Twitter data and data drawn from traditional methods. With this said, the most pressing next steps are:

1. Investigate the underlying causes for the apparent spread of *ain't*+inf south from the northern cities to those in the south, while not to western coastal areas. Could this be explained by the maintenance of cultural ties (or lack thereof) and migration patterns?

2. Investigate other forms of *ain't*: correlations between other forms of *ain't* and BAA Populations in order to check the hypothesis that Twitter data shows that *ain't*+inf is uniquely tied to AAE. For example, do other uses of *ain't*, particularly *ain't+perfective* (as in *ain't seen*), also show a high correlation with BAA communities?

3. Investigate weak verbs: the extent to which weak verbs (*move*, *raise*), consonant cluster reduction (etc.) and other phonological properties of AAE influence the orthographical spelling of verbs on Twitter, and could mean some of what appears to be *ain't*+inf is actually perfective *ain't*.

4. Investigate contextual variation: the extent to which contextual variation exists in the use of *ain't*+inf across different regions.

5. Build the Atlas. Begin the process of compiling all of this into an interactive atlas. Combine different visualisation methods (such as pie charts in first map). Research and add other AAE parts of speech.

6. Investigate alternative methods. Use alternative method for getting location data – *I grew up in* rather than GPS point from where a tweet was sent.

7. Investigate outliers among box plots and scatter plots. Identify the locations of outlier census tracts, in order to further investigate the context behind outlier tracts. In the case of the boxplots, this context also addresses the outliers in the BAA minority group, and the disparity in the number of outliers in the BAA majority group.

Abbreviations

BAA Black and African American
inf infinitive verb

References

Bachelier, Veronique, Jalal-Edine Zawam, Benoit Thieurmel, Francois Guillem & RTE. 2021. *leaflet.minicharts: Mini charts for interactive maps*. DOI: 10.32614/CRAN.package.leaflet.minicharts.

Baxter, Kimberley. 2025. Extracting "non-standard" data from the Twitter API. In Susanne Wagner & Ulrike Stange-Hundsdörfer (eds.), *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII*, 3–30. Berlin: Language Science Press. DOI: 10.5281/zenodo.15006593.

Blodgett, Sue Lin, Lisa J. Green & Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In Jian Su, Kevin Duh & Xavier Carreras (eds.), *Proceedings of the 2016 conference on empirical methods in Natural Language Processing*, 1119–1130. Austin, TX: Association for Computational Linguistics. DOI: 10.18653/v1/D16-1120.

Eisenstein, Jacob. 2013. Phonological factors in social media writing. In Cristian Danescu-Niculescu-Mizil, Atefeh Farzindar, Michael Gamon, Diana Inkpen & Meena Nagarajan (eds.), *Proceedings of the NAACL/HLT 2013 workshop on Language Analysis in Social Media (LASM 2013)*, 11–19. Atlanta, GA: Association for Computational Linguistics. https://aclanthology.org/W13-1102/.

Fisher, Sabriya. 2022. The status of *ain't* in Philadelphia African American English. *Language Variation and Change* 34(1). 1–28. DOI: 10.1017/S0954394522000060.

Gopal, Deepthi, Tamsin Blaxter, David Willis & Adrian Leemann. 2021. Testing models of diffusion of morphosyntactic innovations in Twitter data. In Arne Ziegler, Stefanie Edler & Georg Oberdorfer (eds.), *Urban matters: Current approaches in variationist sociolinguistics* (Studies in Language Variation 27), 253–278. Amsterdam: John Benjamins. DOI: 10.1075/silv.27.

Horvath, Barbara & David Sankoff. 1987. Delimiting the Sydney speech community. *Language in Society* 16(2). 179–204. DOI: 10.1017/S0047404500012252.

Howe, Darin. 2005. Negation in African American Vernacular English. In Yoko Iyeiri (ed.), *Aspects of English negation*, 173–203. Amsterdam: John Benjamins. DOI: 10.1075/z.132.16how.

Jones, Mari C. (ed.). 2015a. *Endangered languages and new technologies*. Cambridge, MA: Cambridge University Press.

Jones, Taylor. 2015b. Toward a description of African American vernacular English dialect regions using "Black Twitter". *American Speech* 90. 403–440. DOI: 10.1215/00031283-3442117.

Jørgensen, Anna, Dirk Hovy & Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In Wei Xu, Bo Han & Alan Ritter (eds.), *Proceedings of the workshop on noisy user-generated text*, 9–18. Beijing, China: Association for Computational Linguistics. DOI: 10.18653/v1/W15-4302.

Kautzsch, Alexander. 2012. *The historical evolution of Earlier African American English: An empirical comparison of early sources*. Berlin: de Gruyter Mouton. DOI: 10.1515/9783110907971.

Labov, William, Paul Cohen, Clarence Robins & John Lewis. 1968. *A study of the non-standard English of Negro and Puerto Rican speakers in New York City*, vol. 2. Philadelphia: U.S. Regional Survey.

Labov, William & Wendell A. Harris. 1986. *De facto segregation of black and white vernaculars* (Current Issues in Linguistic Theory). Amsterdam: John Benjamins. 1–24. DOI: 10.1075/cilt.53.04lab.

Mitchell, Travis. 2019. *Sizing up Twitter users*. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

Moody, Simanique Davette. 2011. *Language contact and regional variation in African American English: A study of Southeast Georgia*. New York: New York University. (Doctoral dissertation).

Nguyen, Dong, Dolf Trieschnigg, A Seza Dog, Mariet Theune, Theo Meder & Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1950–1961.

Orton, Harold & Eugen Dieth. 1962. *Survey of English dialects*. Leeds: E.J. Arnold, University of Leeds.

Rickford, John R., Arnetha Ball, Renee Blake, Raina Jackson & Nomi Martin. 1991. Rappin on the copula coffin: Theoretical and methodological issues in the analysis of copula variation in African-American Vernacular English. *Language Variation and Change* 3(1). 103–132. DOI: 10.1017/S0954394500000466.

Ryan, Camille L. & Kurt Bauman. 2016. *Educational attainment in the United States: 2015. Population characteristics. Current population reports* (Report number: P20-578). Suitland, MD: US Census Bureau. https://www.census.gov/library/publications/2016/demo/p20-578.html.

Stevenson, Jonathan. 2016. *Dialect in digitally mediated written interaction: A survey of the geohistorical distribution of the ditransitive in British English using Twitter*. University of York: University of York. (MA thesis).

Strelluf, Christopher. 2019. Positive-anymore, American regional dialects, and polarity-licensing in tweets. *American Speech* 94(3). 313–351. DOI: 10.1215/00031283-7587883.

Strelluf, Christopher. 2020. Needs+PAST PARTICIPLE in regional Englishes on Twitter. *World Englishes* 39(1). 119–134. DOI: 10.1111/weng.12451.

Tamir, Christine, Abby Budiman, Luis Noe-Bustamante & Lauren Mora. 2021. *Facts about the U.S. Black population*. Washington, DC: Pew Research Center's Social & Demographic Trends Project.

US Census Bureau. 2015. *Data profiles*. https://www.census.gov/programs-surveys/acs/.

US Census Bureau. 2018. *American Community Survey updates: 2018*. https://www.census.gov/programs-surveys/acs/news/updates/2018.html.

Walker, Kyle & Matt Herman. 2023. *tidycensus: Load US census boundary and attribute data as "tidyverse" and "sf"-ready data frames*. DOI: 10.32614/CRAN.package.tidycensus.

Weldon, Tracey. 1994. Variability in negation in African American Vernacular English. *Language Variation and Change* 6(3). 359–397. DOI: 10.1017/S0954394500001721.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo & Hiroaki Yutani. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* 4(43). DOI: 10.21105/joss.01686.

Willis, David. 2020. Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: Two case studies from Welsh. *Glossa: A journal of general linguistics* 5(1). 103. DOI: 10.5334/gjgl.1073.

Wolfram, Walt. 2007. Sociolinguistic folklore in the study of African American English. *Language and Linguistic Compass* 1(4). 292–313. DOI: 10.1111/j.1749-818X.2007.00016.x.

Wolfram, Walt & Natalie Schilling-Estes. 2016. *American English: Dialects and variation*. 3rd edn. (Language in Society 25). Chichester, UK: Wiley Blackwell.

Wolfram, Walt & Erik R. Thomas. 2002. *The development of African American English* (Language in Society 31). Oxford: Blackwell Publishers.