Chapter 1

Extracting "non-standard" data from the Twitter API

Kimberley Baxter
New York University

The present paper examines methodology in the use of Twitter in the corpus-based analysis of African American English (AAE) syntax. I discuss the extraction and geospatial mapping of indices of use of the perfective marker *done* (hereafter *perfective done*), which, alongside a simple past form, indicates that the action described in the past form has been completed. Widespread use of AAE on archived social media posts creates a living database of timed, dated, and geotagged utterances from which corpora may be built. The Academic Twitter API (ACTW) allows access to their full database of tweets, which is a much larger and more accessible dataset than its large social media contemporaries. I discuss two methods of extracting perfective *done* from the ACTW: a *front-end* approach which aims to isolate uses of perfective *done* by eliminating non-perfective uses of *done* from the search prior to running the query, and a *back-end* approach which first extracts a set of all uses of *done* from 2012–2015 and aims to isolate uses of perfective *done* afterwards. I discuss the results of this method, as well as the implications therein and directions for future research. I conclude that while both methods are effective at extracting perfective *done* from the ACTW, the *back-end* approach is better suited to geospatial mapping.

1 Introduction

1.1 General remarks

Despite longstanding myths of African American English (AAE) as a linguistic monolith (Wolfram 2007), numerous sociolinguistic studies indicate that AAE

Kimberley Baxter. 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

shows regional variation in both phonology and lexicon (see Wolfram 2007, Lanehart 2015, etc. for phonology; Jones 2015, Grieve 2016, etc. for lexicon). Far fewer studies focus on syntactic variation in AAE across regions, and of those that do, many are restricted to a certain geographic area (e.g. Terry 2010, Moody 2011), with Baxter (2025 [this volume]), Baxter & Stevenson (2025 [this volume]), and Masis et al. (2022) constituting notable exceptions. This study aims to fill this gap by using geotagged Twitter data to investigate syntactic variation in AAE across the contiguous United States.

The goals of this study are to:

a) Present a method for the targeted search of AAE parts of speech which maximizes the number of desired features while minimizing the number of false positive features, which are orthographically identical to the desired feature, yet still syntactically different;

b) Calculate indices which reflect the rate of feature use by geographical region;

c) Correlate these indices with demographic data to discover whether indices of feature use vary across locations with a high density of Black/African American people.

By doing so, this paper seeks to address the myth of supraregionality in AAE (Wolfram 2007), which suggests nationwide grammatical universality in urban centers (see Section 1.2).

This paper focuses on the methodology used to plot the geographic distribution of one such feature canonical to AAE: perfective *done* (Rickford 1999, Green 2002, 2010). I describe two approaches to the collection of perfective *done* tokens from the Academic Twitter API (ACTW): a *front-end* approach, through which grammatical restrictions are applied prior to extracting tweets, resulting in a dataset comprised mostly of tweets using perfective *done*, and a *back-end* approach, through which grammatical restrictions are applied after extracting tweets, resulting in a dataset comprised of all uses of *done*, from which a subset of perfective *done* may be extracted.

1.2 African American (Vernacular) English

This paper uses *African American English* (AAE) to describe the language previously known as *African American Vernacular English* (AAVE), *Black Vernacular*

English (BEV), etc. In doing so, I aim to distance myself from the "vernacularization" of AAE introduced by previous research which focused mainly on male street youth in the inner city (e.g. Labov 1972). AAE is one of many African American languages which fall under the umbrella of AAL, including, but not limited to, Gullah Geechie, Black American Sign Language, Louisiana Creole, and many more. AAE is a sociolect spoken mostly by Black Americans and people who live in community with Black Americans. In discussing AAE, it is also important to consider race and ethnicity in the United States and how that informs not only what AAE is, but who speaks it, and how African American and other Black people in the US are categorized demographically (see Blake (2014) and King (2020), both of which speak to issues of the racialization and categorization of AAE).

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 utilize the US Census as a touchstone for demographic metadata, including the proposed dissertation. I use *Black/African American* (henceforth: BAA) in discussions regarding the racial and ethnic identity of Black people in this paper.

1.3 Why investigate variation in AAE?

While research on the systematicity of AAE sparked a groundbreaking shift in Sociolinguistics as a whole, researchers also inadvertently established a series of myths about AAE and its uniformity (Wolfram 2007):

(1) The Supraregional Myth: primary structural features setting apart the vernacular speech of African Americans from their European American cohorts were shared by African American Communities regardless of regional context.

(2) The Language Change Myth: a uniform path of change for African American English, based on the uniformity of AAE

(3) The Social Stratification Myth: the prevailing assumption that African American English is most commonly used by working-class speakers, especially those who have had little to no contact with other varieties of English. (Wolfram 2007: 295–306)

These myths persist among the definitions of AAE in previous literature:

> [AAE is] the uniform grammar used by African Americans who have minimal contact with other [varieties] in contexts where only speakers of that vernacular are present. (Baugh 1983, as cited by Labov 1998: 6)
> Due to sociopsychological barriers between Blacks and Whites, AAE is a marker of identity. This causes uniformity among Blacks from all over the US. (Rickford 1999, as cited by Johnson 2008: 20)

The impression of AAE as a "uniform" sociolect is widespread in previous literature on AAE. While these definitions were published decades ago, this myth of uniformity across regions, also known as *supraregionality*, still persists in more recent literature. Wolfram (2007) goes into great detail about all three of these myths, but the most salient of these for the purpose of this paper is the Supraregional Myth. While the papers cited in Wolfram (2007) were from decades earlier, this myth is still very much present in relatively recent literature, as seen below:

> [AAE] is, in contrast to other North American [varieties], not geographically restricted. Although variation in AAE does exist, AAE in urban settings has been established as a uniform system with suprasegmental norms… (Jørgensen et al. 2015: 10)

The use of Twitter as a source of data allows researchers to collect far more data than would be possible via traditional methods such as surveys and sociolinguistic interviews. Twitter's Academic Developer License (ADL) allows access to Twitter's entire archive of tweets, from which a maximum of 10,000,000 tweets may be mined per ADL, per month. Because the archived tweets were produced voluntarily by Twitter users, this method may also circumvent the Observer's Paradox (Labov 1972) and similar issues which often arise during the process of face-to-face data collection. The use of Twitter data also presents a boon for the collection of parts of speech which are more difficult to elicit via interviews because of the presence of linguistic alternatives, through which similar meanings and ideas may be conveyed. For example, where my previous attempts at eliciting perfective *done* from interviews yielded very few results, Twitter allows me to pinpoint parts of speech associated with AAE (Green 2002, Rickford 1999) specifically. This data can then be mapped geospatially via the metadata included with each tweet and provide insight on the geographical distribution of perfective *done*.

2 Analyzing language variation in social media

There are well-documented challenges analyzing non-standard varieties of English commonly used in social media (Plank et al. 2016), and AAE is no different. These challenges fit into two broad categories: grammatical challenges, and demographic challenges. I discuss each of these challenges below.

2.1 Grammatical challenges

Standard methods of extracting data from Twitter's APIs often lack the specifications necessary to isolate parts of speech exclusive to AAE, and differentiate them from similar lexical items in Mainstream American English (MAE) (Jørgensen et al. 2015). This is in part because AAE has a robust verbal system which allows its verbal lexicon to be used as aspect and mood/modality markers in addition to their use as tense markers, as seen below in examples (4), (5), and (6). (4) is an example of perfective *done*), indicating that the action of going to the store has been completed. Example (5) shows perfective *done* being used in conjunction with the simple past form of the verb *do*. Example (6) shows perfective *done* being used in conjunction with *do* in its participle form.

(4) He done went to the store already.
He PERF went to the store already.
'He has gone to the store already.'

(5) He done did his work.
He PERF did his work.
'He has done his work.'

(6) Now look what you done done.
Now look what you PERF PART.
'Now look what you've done.'

All three of these examples are grammatically correct in AAE, yet the perfect and participle forms of *done* in (6) are orthographically identical. For the purpose of this paper, an alternate form of *done* which is grammatically different, yet orthographically identical to perfective *done* is called a *false positive* (FP). Conducting a simple search for the word *done* in the ACTW results in a high number of FPs and data which renders the manual elimination of FPs unfeasible due to the sheer size of the dataset, which numbers in the millions of tweets (see Section 4).

This paper ultimately aims to produce an alternative method which allows the user to eliminate the aforementioned FPs by coding the grammatical constraints of perfective *done*, which would otherwise be inaccessible due to the lack of highly accurate part of speech taggers designed for this task. While the focus of this research is on Twitter, I raise further questions about language variation in AAE in Section 7.

2.2 Demographic challenges

Tracking variation in a sociolect such as AAE is difficult on social media sites which grant anonymity to its millions of users. In the case of 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 completely optional. As a result, a reliable mass verification of users' demographic data is not currently feasible.

Multiple surveys have been conducted in an effort to tease apart the demographics of Twitter, with varying results. This study refers to the 2018 Pew Research Study "Sizing Up Twitter" to describe the broad demographics therein. "Sizing Up Twitter" is a representative survey 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 10% of Twitter 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. Eleven percent of Twitter users identify as Black* (not including Black people who also identify as Hispanic). Of these, approximately 30% are college graduates, 40% have some college education, and 30% have a high school diploma or have not completed their high school education. The survey does not include statistics on political affiliation, age, or use *by race*.

However, the study does present a representative Twitter demographic which shows a percentage of Black tweeters which is comparable to the general population of the United States in the same year (approximately 13% of the total population of the United States; U. S. 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). While there are some gaps between the "Sizing Up Twitter" dataset and the US Census Bureau dataset, the lower percentage of college graduates in the US Census Bureau dataset appears to support the finding that Black Twitter users are also more likely to be college educated than the general population.

With the recent the recent sale and transformation of Twitter (now X) under the ownership of Elon Musk, the demographics therein, especially along racial and political lines, may now be very different. In a Washington Post article entitled "Fleeing Elon Musk's X, the quest to re-create 'Black Twitter'", Dwoskin (2023) describes a political shift in Twitter's formerly left-leaning environment:

> Hate speech has surged on the platform. Researchers with the Network Contagion Research Institute (NCRI), a group that analyzes hundreds of millions of messages across social media, discovered an account that included a Nazi swastika in its profile picture tweeting antisemitic memes. Use of the n-word soared by nearly 500 percent, and the slur popped up in the handle of an account authorized by Musk's subscription service, Twitter Blue. And thus the exodus of Black users began. (Dwoskin 2023)

The data used in both this paper and Baxter & Stevenson (2025 [this volume]) was produced by its users and collected from the ACTW prior to many of these changes, and are reflective of the Twitter captured in Wojcik & Hughes (2019). In addition, the talk from which this paper is derived was given before Twitter became X. For the sake of consistency, this paper uses Twitter to refer to the social media platform now known as X. In addition, there does not appear to be a new term for posts made on the platform; as a result, this paper also maintains *tweet*(*s*) as the label for posts made therein.

Since this talk was given, the ACTW has been discontinued, or at the very least paused, preventing the collection of further data from this source. However, it is my belief that these methods will still prove useful to researchers who wish to extract "non-standard" data from similar social media APIs, including X's new Enterprise API product.

2.3 Linguistic grouping approach

Because of the relative anonymity of Twitter profiles, this study uses a "linguistic grouping" approach (Horvath & Sankoff 1987), through which linguistic features of AAE are collected and analyzed *before* the categorization and analysis of sociological factors such as race and ethnicity. This method is preferred to the traditional "sociological grouping" approach used in most traditional sociolinguistic studies which, because of the aforementioned anonymity of Twitter profiles, makes it difficult to verify the ethnicities of the users behind each account.

> 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 perfective *done*, and calculating indices of use, 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 contiguous United States. By doing so, I aim to provide a broad yet comprehensive view of perfective *done* usage rates in high-density BAA communities.

I plan to conduct future research (see Section 7) to properly investigate language differences among first- and second-generation Black immigrants' AAE and the AAE spoken by Black people who are ethnically African American. However, this is beyond the scope of this paper, as well as the limits of what Twitter data or US Census data can currently offer.

3 Perfective *done*

Similarly to Blodgett et al. (2016), Stevenson (2016), and Willis (2020), the present study aims to examine syntactic variables in AAE. This should not be confused with studies on the spread and use of lexical items in AAE such as *fleek* (Grieve 2016), *eem* (Jones 2015), or others. This distinction is very important because AAE has a robust verbal system which allows the use of verbal items as tense, mood/modality, and aspect markers. It is not enough to simply search the ACTW for all uses of *done*, *be*, *BIN* (spelled *been* by AAE speakers), or any other part of AAE speech. One must be familiar with the syntactic properties of these items and devise alternate strategies to separate them from their verbal, adjectival, or other counterparts when extracting data from the ACTW. It is with that in mind that I discuss perfective *done*.

Green (2002) describes perfective *done* below. Numbers have been changed for continuity within the present paper:

(7) a. I told him you dən changed. (Bm, 30's)
'I told him that you have changed.'
A: You through with Michael Jordan I bought you?
(Literally: Have you finished reading the magazine that I bought you with Michael Jordan on the cover?)
B: I dən already finished that. (Bm, 9)
'I have already finished that.'

b. I dən done all you told me to do. I dən visited the sick. (Bm, 60s, 70s)
'I have done all you told me to do. I have visited the sick.'

c. A: Push your seat.
B: I dən pushed it.
'I have (already) pushed it.'
A: Push it again. (elderly Bfs on Amtrak)
(Green 2002: 60)

(8) a. My homework is *done*.
b. Have you *done* your taxes?

Perfective *done* is distinct from *done* in its verbal or adjectival form as seen above in example (8) and is most often compared to the MAE perfect markers *have*, *has*, and *had*.

Perfective *done* typically occurs with both stative and eventive verbs, as well as time adverbs and negation (Martin 2018). The present study uses the descriptor *perfective* rather than *completive* because *perfective* is inclusive of uses of *done* which indicate that an action, event, or status has been completed, as well as uses of *done* which indicate that an action, event, or status is still ongoing.

Due to its ability to describe completed actions, events, or statuses, perfective *done* often appears with eventive verbs, which have a natural endpoint. However, in certain contexts, perfective *done* can occur with stative verbs, which do not have natural endpoints. For example, the following sentence includes perfective *done* with the stative verb *know*:

(9) Long as I done known you (you've been chasing after women)
TMP-long as I PERF known you (you've been chasing after women)
'For as long as I have known you (you've been chasing after women).'
(Wilson (1986), cited in Martin (2018))

Terry (2010) argues that *done* is perfective because it fits the "four perfect constructions" as outlined in Comrie (1976):

(10) Perfect of persistent situation: this perfect behaves similarly to stative BIN, in that it indicates that an event began in the past, and is still occurring.
"Mary *done* lived in Chapel Hill for 6 years."
The above example indicates that Mary has been living in Chapel Hill since the distant past, and still does.

(11) Experiential perfect: indicates that a situation took place or held at some time in the past.
"They *done* took my car."

(12) Perfect of result: the present state is referred to as being the result of the past.
"And now you *done* messed everything up."

(13) Perfect of recent past: temporal closeness to the present is the focus, rather than the completion.
"John *done* just got here."

Terry (2010) notes that perfective *done* expresses the "continuing relevance of a previous situation" (Comrie 1976: 56) and, as a result, appears to fit the criteria of a perfect marker.

However, it is important to note that while Terry (2010) makes a convincing argument as to why *done* should be a perfective, the use of the "four perfect" structure in Comrie (1976) does not adequately describe perfective *done* in AAE. Certain uses of perfective *done* do not fit neatly into these categories and can even occupy multiple categories at once. For example, when asking informally about the categorization of the utterance listed in example (9), *Mary done lived in Chapel Hill for 6 years*, several AAE speakers agreed with the judgment that it meant *Mary* still lived there while others disagreed, stating that *Mary* lived there already and now lived somewhere else. This difference in interpretation may be the result of several factors including context taken from voice intonation, context from previous portions of the conversation, or in some cases a lack of context as a result of the utterance being written via text or survey.

In addition, while perfective *done* can be compared to use of the perfect in MAE (*have*, *has*, *had*), it is not grammatically *identical* to the perfect in MAE (Martin 2018).

Perfective *done* also appears with adverbs that refer to a time in the past, such as *yesterday* in (14a) below. This is in contrast to the standard English perfect, which cannot appear with such adverbs, as illustrated in (14b). Numbers have been changed for continuity with the present paper.

(14) a. John done baked a cake yesterday. (AAE)
b. * John has baked a cake yesterday. (standard English)

This may be surprising given that the standard English perfect is otherwise quite similar in meaning to perfective *done* (Martin 2018).

To my knowledge, there is no research that specifically examines perfective *done* in Northern cities such as New York City, but there is a wealth of knowledge on perfective *done* in other regions, particularly the Southern United States (e.g. Terry 2010, Green & Roeper 2007).

I discuss two methods of extracting perfective *done* from the ACTW: a *front-end* approach which aims to isolate uses of perfective *done* by using syntactic parameters to eliminate non-perfective uses of *done* from the search prior to extracting data from the ACTW, and a *back-end* approach which first extracts a set of all uses of *done* from the ACTW and isolates uses of perfective *done* from the resulting dataset. I discuss the results of each method, as well as the implications therein and directions for future research. I conclude that while both methods are effective at extracting perfective *done* from the ACTW, the *back-end* approach is better suited to geospatial mapping.

4 Methods

4.1 The front-end approach

To extract data from the ACTW, one must first tell the ACTW exactly what items to extract in the first place. However, as mentioned above, if the ACTW is told simply to extract uses of the word *done*, it will extract a sample of all uses of *done* within the requested time frame (2012–2015) with no distinction between tense, aspect, or other grammatical forms. The resulting dataset would be a total of some 8.5 million tweets taken from approximately 1.6 million individual users. Where Willis (2020) was able to eliminate irrelevant users and their respective tweets manually from a relatively small dataset, this becomes unfeasible with such a relatively high number of tweets. In addition, existing part of speech taggers often do not tag AAE parts of speech accurately (Jørgensen et al. 2015, Blodgett et al. 2016). This issue is further complicated by the fact that the ACTW also does not adhere to punctuation in the separation of multiple utterances within the same tweet. Add to this the fact that perfective *done* is orthographically identical to other *done* forms, and it becomes clear that alternate methods must be used to extract perfective *done* in a way that includes as few "don't count cases" (Blake 1997) as possible.

For the *front-end* approach, which aims to extract a perfective *done* dataset directly from the ACTW, I first establish the syntactic parameters within which perfective *done* occurs and, most importantly, the parameters through which perfective *done* does not occur. I include a summary of "count" and "don't-count"

cases below to illustrate which orthographical instances of *done* were included and which ones were not. See Table 1.

Table 1: A list of *done* uses, their properties, and examples of each alongside their status as included (yes) or not included (no)

Syntactic Parameter	Example	Perfective or Verbal	Included?
a. *done* + participle	"He *done* gone already."	Perfective	Yes
b. *done* + simple past	"He *done* went already."	Perfective	Yes
c. *be* + *done*	"By the time they finish he('ll) be *done* left already."	Perfective (future)	No
d. conjugated *be* + *done*	*is done, are done, am done, was done, etc.*	Verbal* (see: contracted copula)	No
e. *done* + progressive	"She's *done* working."	Adjective	No
f. *done* + preposition	"Wait till the work is *done* to start packing everything up."	Adjective	No
g. phrase-final *done*	"The laundry is done. Sam will take care of the rest."	Adjective	No
h. *done* + contracted copula	"I'm done lost my mind!"	Perfective	No

4.1.1 Simple past and participle

As mentioned above, most cases of perfective *done* are *done* + simple past, with *-ed* endings or *done* + participles (Green 2002, Martin 2018). Most importantly, when plugged into the ACTW, these forms yield the greatest number of positive matches to the desired syntactic variable.

4.1.2 Be done

While *be done* is a form of perfective *done*, preliminary searches revealed a high number of false positives among the results, as seen below in Example (15):

(15) a. I'll be done in a minute.
b. By then the job will be done. Left the door open for you.

Example (15a) is an example of a false positive garnered by searching *be done* only. Example (15b) is an example of a false positive garnered by searching *be done left*. While (15a) may be resolved by adding a past tense verb to the *be done* construction, (15b) shows an example of a false positive which appears even with the addition of a past tense verb. This is because the ACTW ignores punctuation in its searches. Because of the high numbers of false positives among the *be done* dataset, I exclude it from this study.

4.1.3 Conjugated copula and auxiliary *be*

Perfective *done* does not typically occur with full forms of conjugated copula and auxiliary *be*, as seen in (16a). This is not to be confused with the contracted copula, as seen in (16c).

(16) a. * I am done left already.
b. ✓ I done left already.
c. ✓ I'm done left already.

The other cases listed above in Table 1 are representative of verbal or adjectival forms of *done*. These items were eliminated because they do not fit the syntactic parameters of perfective *done*. The search is conducted in a very similar way to the way in which the search for *ain't-for-didn't* (Baxter & Stevenson 2025 [this volume]) is conducted, as shown below in Figure 1.

Figure 1 is an image of the for loop used to collect data from the ACTW for *ain't-for-didn't*. In the case of perfective *done*, I search for *done* + a list of approximately 150 commonly used verbs in their simple past and participle forms, while excluding the aforementioned "don't count" (Blake 1997) cases. The resulting dataset extracted directly from ACTW was approximately 526,000 tweets.

I then examined the resulting data to check for FPs. I then arranged the tweets by "place name," according to the metadata within each tweet. The place name is represented as "City, State" or "City, Territory" within the metadata, and represents the geographical location where the user was when they published the tweet. I first examine the output of a large city with between 5 and 10 thousand tweets (such as Charlotte) and then I document the "don't count" and other cases therein which show high numbers of FPs. Some examples of constructions with a with high numbers of FPs are listed below. Queries are italicized, FPs are marked with a star*, and examples of perfective *done* are marked with a check ✓.

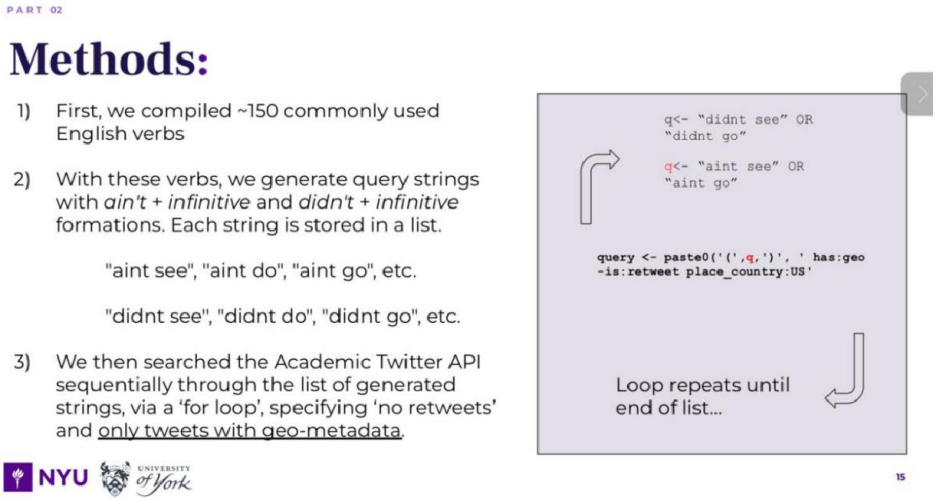

Figure 1: Baxter & Stevenson (2025 [this volume]), slide 15; screenshot of loop for collecting tweets from the ACTW

(17) *I've done drunk*

a. * I climbed onto a roof last night. Probably the #1 unsafe thing I've done drunk

b. ✓ I've done drunk my problems away before

(18) *done done*

a. * Done Donee done done done done done done.

b. ✓ I done done a lot in my 21.9 years of my life

(19) *done did*

a. * Done done done. Did I mention I was done?

b. ✓ And I done did everything but trust these hoes

The most common FPs included *be done, done done, have done*, contracted-copula *done*, and *done*-adverb constructions, including temporal adverbs such as *just*, *already*, and *yesterday*. All of these can be used in perfective *done* constructions (see Table 1); however, as mentioned above, due to the large number of FPs within the query results and/or the relative rarity of these cases among the data, they were eliminated.

After eliminating strings which were most likely to lead to FPs, 426,000 tweets of perfective *done* remained. This set of tweets will hereafter be called the perfective *done* dataset.

I calculate an index of perfective *done* use in a similar way to the "straight" model outlined in (Rickford et al. 1991). As a result, it is necessary to find something to which perfective *done* may be compared. As stated above, while perfective *done* can be compared to the MAE perfect (*has*/*had*/*have*), it is not identical. In addition, because it is not identical to the MAE perfect and because of the comparative robustness of the verbal TMA system in AAE, there is no single one-to-one yardstick to which perfective *done* may be compared. In an attempt to solve this problem, I choose a calculation in which the *perfective done dataset* is divided by all uses of *done* taken from the same time period (2012–2015); this sample of 8.5 million tweets will hereafter be called the *all done dataset*.

Both the *perfective done dataset* and the *all done dataset* are cleaned in several ways. First, all duplicate tweets are removed from each file to eliminate mass-duplicate tweets containing song lyrics and popular turns of phrase which may not be used in everyday speech. In addition, all utterances with three or more consecutive instances of *done* are removed to eliminate utterances like the one in Example (18a), which are not examples of perfective *done*. Each dataset is then reduced to one tweet per user. This helps to eliminate the misrepresentation of tweets by individual users who tweet more heavily than others. I choose this method rather than the usual method of gathering proportions for each individual because this study frames perfective *done* use in terms of how many twitter users in each location use perfective *done*, rather than how many times each user uses perfective *done,* or what proportion of use each user produces. Limiting tweets to one per user is a clear and straight-forward way to figure out how many twitter users produced the token within a given area.

Once each dataset is reduced to one-per-author, I extract the latitude, longitude, place name, and user id, and calculate a per-city count based on the number of tweets occurring within each location. For the *perfective done dataset*, this count will be called the *perfective done count*, and for the *all done dataset*, this column will be called the *all done count*.

I divide the perfective *done* count by the *all done* count, and the result is a series of indices reflective of the number of users who have used perfective *done* with the selected 150 words in their tweets.

These indices (marked "newindex" on the right side of Figure 2) are mapped according to their geotag and assigned a color gradient, from dark blue for low indices and yellow for high indices. The resulting map is pictured below.

While the *front-end* approach is effective for extracting perfective *done* directly from the ACTW, it is not suitable for calculating indices of use for the following reasons:

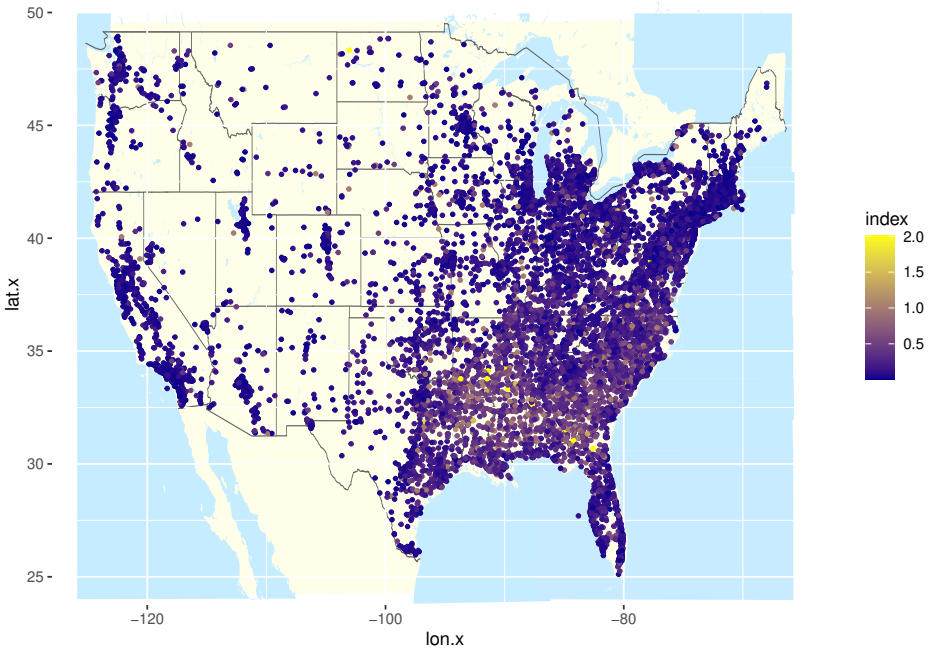

Figure 2: Front-End perfective *done* map

I. The data extracted from the ACTW is not an exhaustive account of all tweets from that time period (https://developer.twitter.com). Rather, it is a collection of data from a given time period which, while still a good sample of Twitter use, may still include or exclude tweets from the *all done dataset*, which was pulled from the ACTW separately, and is thus a separate sample. As a result, indices of certain towns and cities, especially those with a total of 30 user IDs or less, exhibit indices larger than 1. This should not be possible, since all uses of perfective *done* should occur within the *all done dataset*.

II. In restricting the *all done dataset* to one tweet per user ID, I did not account for the fact that the *all done dataset* included *perfective done dataset* tweets among its 8.5 million tweets. If a user appearing in the *all done dataset* used both perfective *done* and verbal or other forms of *done*, there is no way of knowing which tweet R would use to represent that user. As a result, the resulting 1.6 million tweets, restricted to one per user may not be representative of the *unrestricted all done dataset*.

III. When comparing the *perfective done dataset* to the *all done dataset*, it quickly became apparent that the *perfective done dataset* was missing a great many instances of perfective *done* that still appeared in the *all done dataset*. This is because this method specifically searched for *done* + ~150 commonly used verbs. Those instances of *done* were coded as perfective *done*, while all instances of *done* + other verbs were not coded and therefore not included in the *perfective done dataset*, or coded as perfective *done* within the *all done dataset*.

As a result of these and other problems, I employ a different strategy to formulate the *perfective done dataset* and calculate indices of perfective *done* use.

4.2 The back-end approach

Having already formulated the *all done dataset* by extracting all uses of *done* from the ACTW, I extract a new *perfective done dataset* from the existing *all done dataset*, rather than extracting the *perfective done dataset* directly from the ACTW. This immediately remedies the issue of mismatching Twitter samples and by proxy eliminates the occurrence of indices greater than 1.

To do so, I create a list of all lemmas following *done* among the 8.5 million tweets in the *all done dataset*.

While reading all 8.5 million tweets is unfeasible within the given timespan of this study, there are only about 53,000 lemmas following *done* in all 8.5 million tweets, including all alternate spellings of each word (e.g.: *left*, *leff*, *lef*, etc.). I code all ~53,000 lemmas manually and mark simple past forms, participle forms, and any alternate spellings therein with an *x*. This takes roughly 18 hours, but is an incredibly important step in the creation of a database of alternate spellings of past-tense verbs, many of which may not have been predicted by simply guessing what ways users can and cannot spell things.

In addition to the "don't count" (Blake 1997) cases above, I also do not mark *un*-verbs, and other lemmas which can also be interpreted as adjectives because these are also likely to be FPs (e.g.: adjectives such as *undone*, *unloved*, etc.) I then code the resulting ~3,900 marked tokens as perfective *done* and all unmarked tokens as *other*. I then extract all perfective *done* coded tweets to create the new *perfective done dataset*, and extract all *other* coded tweets to a new dataset called the *other dataset*.

Similarly to the *front-end* method above, I restrict each file to one per user, and count the number of instances by city. I then calculate indices with the equation in example (20).

12344	34128	porterhouse	4	other
12345	34140	portraits	4	other
12346	34148	pose	4	other
12347	34149	posed	4	Pdone
12348	34162	posse	4	other
12349	34165	posessed	4	Pdone

Figure 3: Sample image of perfective *done* documentation

(20)

$$\text{perfective } \textit{done} \text{ index} = \frac{\text{perfective } \textit{done} \text{ count}}{\text{perfective } \textit{done} \text{ count + other count}}$$

This new calculation of the perfective *done* index addresses the user representation problem in the *front-end* method, as speakers who use both perfective *done* and other forms of *done* now have both uses adequately represented in the denominator. Once the indices are calculated, I use geospatial mapping tools to map these indices across the contiguous United States as shown below in Section 5.

5 Results and analysis

Figure 4 is the map resulting from the *back-end* approach mentioned above in Section 4.2. Bright (yellow) dots are representative of high perfective *done* use, while dark (blue) dots are representative of low perfective *done* use. A cluster of bright dots can be seen in the southeastern United States, which indicates higher indices of use in this region than in northern, or midwestern regions of the United States. Based on this, the data shows that perfective *done* is more commonly used among Twitter users in the southeastern United States than in other regions.

To confirm these results, I cross-reference these perfective *done* indices with BAA population rates in US Census Tracts in a selection of states representative of each region: Pennsylvania, New York, Illinois, Alabama, and Georgia. Each scatterplot is cropped down to their maximum perfective *done* and BAA population indices. Where most states have some BAA populations that approach 100%, thereby necessitating the need for the full 1.0 on the y-axis, the maximum indices on the x-axis vary widely. For example in Figure 7 (Georgia), the maximum perfective *done* index is approximately 0.7, whereas in Figure 6 (New York), the maximum perfective *done* index is 0.3. This is done for the sake of visibility in

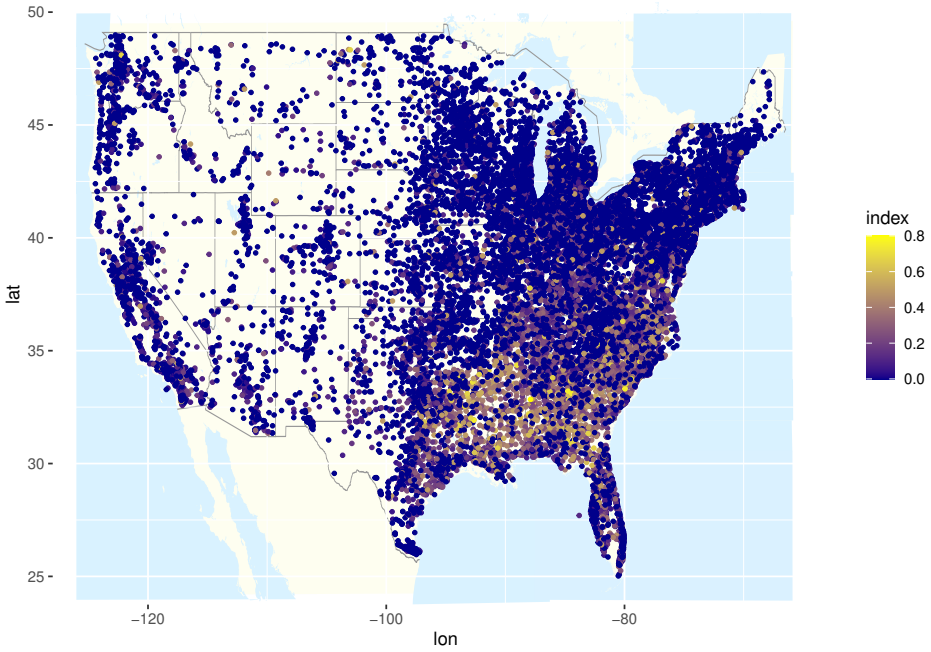

Figure 4: Perfective *done* use in the contiguous United States

states where perfective *done* indices are relatively low, so that the distribution of datapoints therein are still visible and interpretable.

Figure 5 is a scatter plot depicting a comparison of the BAA index: the percentage of BAA-identifying people living in a given area according to the US Census, to the perfective *done* index: the percentage of users with perfective *done* in their tweets. In Pennsylvania, among areas with high populations of BAA-identifying people, indices are relatively low, falling entirely below 0.2.

Figure 6 shows a similar trend to the one shown in Figure 5, with the maximum perfective *done* index at just over 0.3 for any recorded locations in the state. Similarly to Pennsylvania, locations in New York with high populations of BAA-identifying people mostly show perfective *done* indices of 0.2 or less, although very few are just above 0.2. Most notably, New York appears to show many locations with very low perfective *done* indices. Many of these are flatly at the 0.00 mark.

Figure 7 is a scatter plot from Illinois, which shows similar features to previous northern states. Similarly to New York, the maximum perfective *done* index is 0.4. However, unlike New York, Illinois does not appear to show many locations with

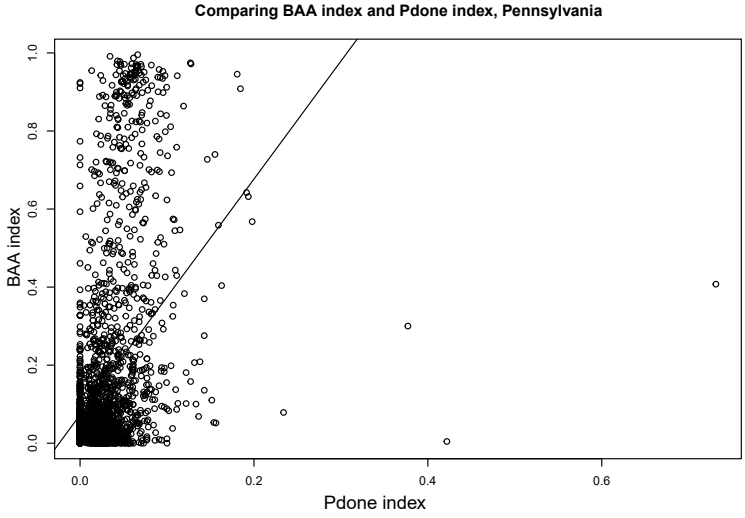

Figure 5: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of Perfective *done* use (Pennsylvania; PDONE MAX VALUE = approx. 0.8).

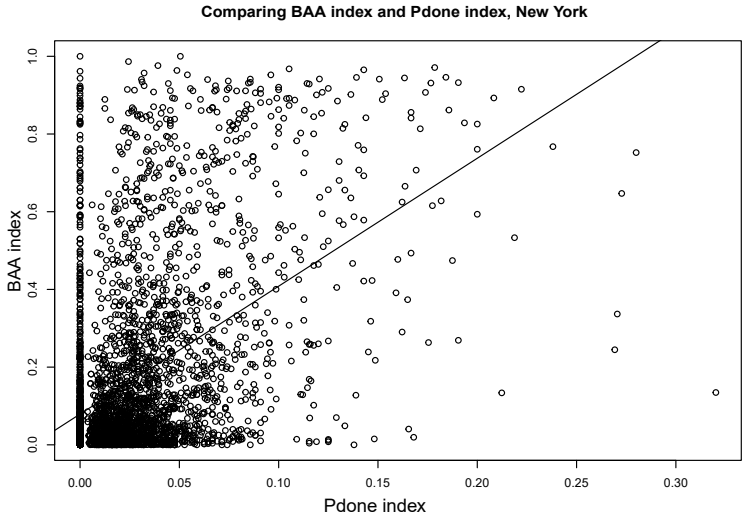

Figure 6: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of Perfective *done* use. (New York; PDONE MAX VALUE = approx. 0.35)

zero ratings. Rather, similarly to Pennsylvania, many locations with high BAA populations appear to be spread between the 0.0 and 0.2 mark, with several more between the 0.2–0.3 marks. While Illinois is a northern state, it is also located in the Midwest. The aforementioned divergence from northeastern states like New York whereby relatively few high-density BAA census tracts in Illinois show a 0.0 perfective *done* index may be indicative of a broader trend in the Midwest. Conversely, it is also possible that New York is unique in its apparent rejection of perfective *done* among Twitter users.

Figure 8 is a scatter plot of data taken from North Carolina. Where many locations with high BAA populations in Northern states appear to max out at between 0.2 or 0.3 perfective *done* index, high BAA populations in North Carolina are mostly between 0.1 and 0.4, with none at the 0.00 mark.

Figure 9 is a scatter plot of data taken from Georgia. Similarly to North Carolina, while there is one location with a high BAA population with a perfective *done* index approaching 0.00, most others fall above 0.1, and show much wider distribution than Northern states, with indices ranging up to 0.5, which is higher than the maximum perfective *done* index for all Northern states mentioned above except Pennsylvania, which has one outlier above 0.6.

Figure 10 is a scatter plot of data taken from Alabama. Similarly to Georgia and North Carolina, while there are two high BAA population locations which fall below the 0.1 mark, most others are diffused along a wider range from 0.1 to 0.5.

Scatterplots from northern states in the eastern and midwestern United States show a greater number of areas with high density populations of BAA people with lower indices of perfective *done*. This indicates a lower rate of use, on average, of perfective *done* among Twitter users in these regions. In contrast, scatterplots from Southeastern states show very few locations with high-density populations of BAA people who do not produce perfective *done*. As indicated in Figures 8–10, maximum indices are higher, and the top left corner is empty or mostly empty, which indicates a lack of locations with high-density BAA populations and low perfective *done* indices. Scatterplots of data taken from southeastern states also shows a wider distribution of perfective *done* than northern cities.

These results confirm the hypothesis that regional differences in indices of perfective *done* use will be reflected in the geospatial data, as shown in maps produced by both the *front-end* and *back-end* methods, as well as the scatterplots above. This also confirms the hypothesis that use of the perfect marker *done* is more concentrated in the southeastern states than in northern states.

Additionally, variation in AAE can be mapped geospatially via the geolocation data included in tweets mined from the ACTW.

Figure 7: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of perfective *done* use. (Illinois; PDONE MAX = approx. 0.4)

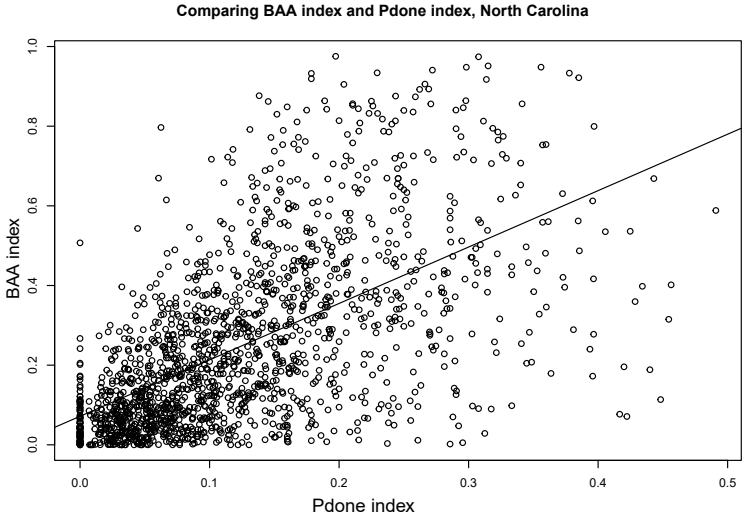

Figure 8: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of Perfective *done* use). (North Carolina, PDONE MAX VALUE = approx. 0.5)

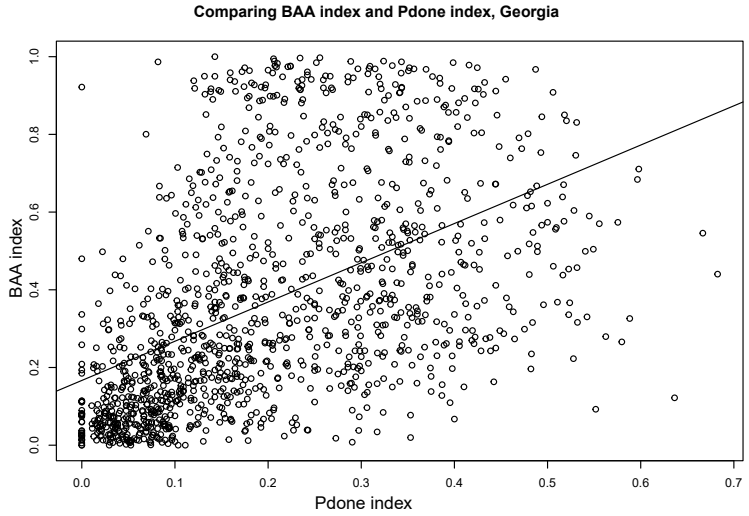

Figure 9: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of perfective *done* use). (Georgia; PDONE MAX VALUE = approx. 0.7)

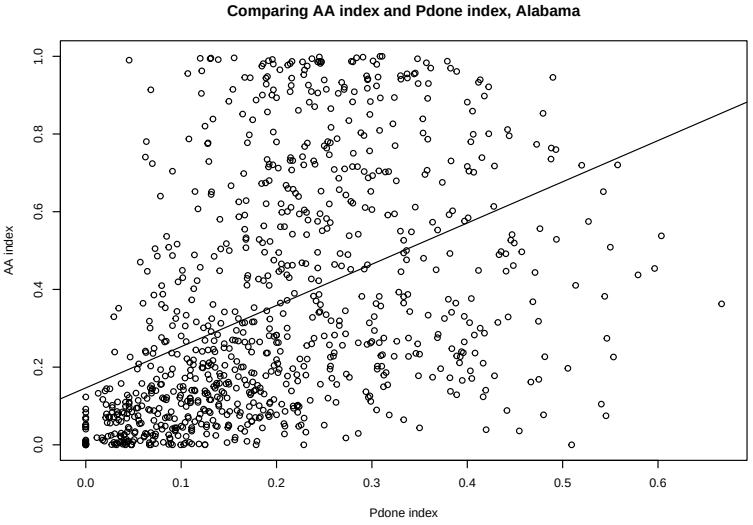

Figure 10: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of Perfective *done* use). (Alabama; PDONE MAX = approx. 0.7)

6 Conclusion

This paper discusses the extraction of perfective *done* and the geospatial mapping of indices of perfective *done* use via the location data attached to each tweet. I discuss two methods of extracting perfective *done* from the ACTW: a *front-end* approach which aims to isolate uses of perfective *done* by eliminating non-perfective uses of *done* from the search prior to running the query, and a *back-end* approach which first extracts a set of all uses of *done* from 2012–2015 and aims to isolate uses of perfective *done* afterwards. I conclude that while both methods are effective at extracting perfective *done* from the ACTW, the *back-end* approach is better suited to geospatial mapping.

The resulting map map shows a cluster of high indices of use in the southeastern United States, which indicates a higher concentration of perfective *done* use among Twitter users therein. Indices of use of perfective *done* were then cross-referenced with demographic data from the US Census. The resulting scatter plots showed that overall, Census tracts with high-density populations of BAA people in New York and Pennsylvania had much lower indices of perfective *done* use than similar populations in Alabama, Georgia, and North Carolina. Not only does this challenge previous assumptions of a "uniform" AAE in urban centers across the United States, it indicates a likely correlation between region and perfective *done* use, with northeastern states showing lower indices overall, and southeastern states showing higher indices overall. Illinois presents an interesting case wherein indices of perfective *done* use in high-density BAA populations are slightly higher than northeastern states, but still lower than southeastern states. It may be that this middle-ground result is indicative of data taken from midwestern states – however, further study must be conducted regarding indices of use in other midwestern states.

Finally, I have shown that a language grouping approach is an appropriate method in the case of anonymized social media data which presents a variety of challenges in identifying and classifying the ethnicities of users. By using this approach and using demographic data taken from the US Census, I provide a broad yet comprehensive view of perfective *done* usage rates in high-density BAA communities across the United States.

7 Future research

In the future, I plan to add additional methods to this study to further legitimize the use of social media data and Census data to describe patterns in AAE on such a wide scale. I outline the next steps in this research below:

1. Investigate the new demographics of X (formerly Twitter), and conduct a deeper examination of how recent leadership, moderation, and other administrative changes have affected Black Twitter. Has the hostile environment outlined in Dwoskin (2023) changed the way language is used among those who continue to use X?

2. Revisit the distribution of perfective *done* in northern states and cities. Do BAA populations in New York, Pennsylvania, and other northeastern or mid-Atlantic states all show low indices of perfective *done* use? Are there pockets of high use in any of these states?

3. Further examine perfective *done* use in midwestern states.

4. Expand methodology to include sociolinguistic surveys and interviews, which are better suited toward topics around sociological grouping, such as racial and ethnic variation in AAE use among L1 speakers.

5. Investigate semantic and pragmatic variation in how perfective *done* is used by the people who use it. While the categories in Comrie (1976) fell short of encapsulating all potential uses of perfective *done* in AAE, investigating distinctions between different types of perfective *done* use is still valuable to our knowledge of how AAE works, how AAE is used, and by whom.

6. Investigate other parts of AAE speech to further deepen our understanding of language variation and change therein.

Abbreviations

ADL	Academic Developer License
ACTW	Academic Twitter API
AAE	African American English
BAA	Black/African American
FPs	False Positives
MAE	Mainstream American English
PDONE	Perfective *done*
PERF	Perfect marker

References

Baugh, John. 1983. *Black street speech: Its history, structure, and survival*. Austin: University of Texas Press.

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.

Baxter, Kimberley & 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.

Blake, Renée. 1997. Defining the envelope of linguistic variation: The case of "don't count" forms in the copula analysis of African American vernacular English. *Language Variation and Change* 9(1). 57–79.

Blake, Renée. 2014. African American and black as demographic codes. *Language and Linguistics Compass* 8(11). 548–563.

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.

Comrie, Bernard. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*, vol. 2. Cambridge: Cambridge University Press.

Dwoskin, Elizabeth. 2023. Fleeing Elon Musk's "X": The quest to recreate "Black Twitter". *The Washington Post* August 6, 2023. https://www.washingtonpost.com/technology/2023/08/06/musk-black-twitter-spill/.

Green, Lisa J. 2002. *African American English: A linguistic introduction*. Cambridge: Cambridge University Press.

Green, Lisa J. 2010. *Language and the African American child*. Cambridge: Cambridge University Press.

Green, Lisa J. & Thomas Roeper. 2007. The acquisition path for tense-aspect. Remote past and habitual in child African American English. *Language Acquisition* 14(3). 269–313.

Grieve, Jack. 2016. *Regional variation in written American English*. Cambridge: Cambridge University Press.

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

Johnson, Sasha Rosena. 2008. *Acknowledging the voices of families: Metadiscourse and linguistic identity of African American speakers of AAE*. University of Georgia. (Doctoral dissertation).

Jones, Taylor. 2015. 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.

King, Sharese. 2020. From African American vernacular English to African American language: Rethinking the study of race and language in African Americans' speech. *Annual Review of Linguistics* 6. 285–300.

Labov, William. 1972. Some principles of linguistic methodology. *Language in society* 1(1). 97–120.

Labov, William. 1998. Coexistent systems in African-American vernacular English. In Salikoko S. Mufwene, John R. Rickford, Guy Bailey & John Baugh (eds.), *African-American English: Structure, history and use*, 154–200. New York: Routledge.

Lanehart, Sonja (ed.). 2015. *The Oxford handbook of African American language*. New York: Oxford University Press.

Martin, Katie. 2018. Perfective *done*. In Raffaella Zanuttini, Laurence Horn & Jim Wood (eds.), *Yale grammatical diversity project: English in North America*. New Haven, CT: Yale University. https://ygdp.yale.edu/phenomena/perfective-done.

Masis, Tessa, Anissa Neal, Lisa J. Green & Brendan O'Connor. 2022. Corpus-guided contrast sets for morphosyntactic feature detection in low-resource English varieties. In Oleg Serikov, Ekaterina Voloshina, Anna Postnikova, Elena Klyachko, Ekaterina Neminova, Ekaterina Vylomova, Tatiana Shavrina, Eric Le Ferrand, Valentin Malykh, Francis Tyers, Timofey Arkhangelskiy, Vladislav Mikhailov & Alena Fenogenova (eds.), *Proceedings of the first workshop on NLP applications to field linguistics*, 11–25. Gyeongju, Republic of Korea: International Conference on Computational Linguistics. https://aclanthology.org/2022.fieldmatters-1.2.

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).

Plank, Barbara, Anders Søgaard & Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics*, 412–418. Berlin, Germany: Association for Computational Linguistics.

Rickford, John R. 1999. Phonological and grammatical features of African American Vernacular (AAVE). In John Rickford (ed.), *African American vernacular English: Features, evolution, educational implications*, 3–14. Malden, MA: Blackwell.

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).

Terry, J. Michael. 2010. Variation in the interpretation and use of the African American English preverbal done construction. *American Speech* 85(1). 3–32.

U. S. Census Bureau. 2018. *2015 annual social and economic supplement: Current population survey*. https://www.census.gov/data/datasets/2015/demo/cps/cps-asec-2015.html. Accessed: 18.09.2023.

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.

Wilson, August. 1986. *Fences*. New York: Plume.

Wojcik, Stefan & Adam Hughes. 2019. *Sizing up Twitter users*. Washington, D.C.: PEW Research Center. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

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.