Now We Stronger Than Ever: African-American Syntax in Twitter

Ian Stewart
Dartmouth College
Hanover, NH 03755
ian.b.stewart.14@dartmouth.edu

Abstract

African American English (AAE) is a well-established dialect that exhibits a distinctive syntax, including constructions like habitual *be*. Using data mined from the social media service Twitter, the proposed senior thesis project intends to study the demographic distribution of a subset of AAE syntactic constructions. This study expands on previous sociolinguistic Twitter work (Eisenstein et al., 2011) by adding part-of-speech tags to the data, thus enabling detection of short-range syntactic features. Through an analysis of ethnic and gender data associated with AAE tweets, this project will provide a more accurate description of the dialect's speakers and distribution.

1 Introduction

Most modern studies of sociolinguistics focus on phonetic or lexical variation to draw conclusions about a dialect or a social group. For example, the Atlas of North American English (2005) maps language variation entirely by the differences in production and perception of phonetic variables. Although this is an integral part of sociolinguistics, research has given less attention to synchronic variation in syntax, which is also an important aspect of language change. Recent initiatives like Yale's Grammatical Diversity Project (2014) have been invaluable in demonstrating the breadth of syntactic variation in North America, and smaller-scale research like Kendall et al. (2011) has been equally vital for investigating the properties of constructions within a "nonstandard" dialect. While other sociolinguistic studies have used a systematic analysis of corpora to detect phonetic and lexical change (Yaeger-Dror and Thomas, 2010; Eisenstein et al., 2011), such approaches are under-utilized with respect to syntactic variation.

Varieties of African American English provide a wide range of syntactic features to study, with constructions ranging from aspectual particles like *done* (such as "he done eaten" for "he's just eaten") to double negation (such as "can't nobody") (Wolfram, 2004). AAE shares some features with Southern American English but is spoken throughout the United States. The majority of research in AAE syntax relies on data collected from interview-based conversations (Labov, 2012), published letters (Kendall et al., 2011) and observations of dialect acquisition in children (Green and Roeper, 2007). Though valuable, this kind of data is often restricted to a specific location and cannot always keep pace with the most recent language developments among fluent young speakers. The proposed study seeks to systematically study AAE syntax in a more youth-centric environment and describe the geographical or gender-based correlation in the distribution of such syntax.

2 Proposal

This thesis's primary hypothesis is that there is a quantifiable correlation between ethnicity and features of AAE syntax found in large-scale social media. This will be supported or challenged by the geographic and demographic data associated with the constructions, as previous studies of dialect reappropriation have suggested a spread of AAE beyond expected areas (Reyes, 2005). As a secondary hypothesis, the project will investigate a correlation between AAE syntax and gender, which has been suggested but not tested on a large scale. Eckert and McConnell-Ginet (2013) argue for a connection between gender and identity expression (often associated with "speech style"), which would generally suggest greater AAE syntax usage among women. Even if the neither correlation is proven plausible, the study will provide valuable insight about the frequency and ge-

Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 31–37,
Gothenburg, Sweden, April 26-30 2014.

ographic location of specific AAE syntactic features. This project is being co-supervised by a professor of sociolinguistics and a postdoctoral researcher in computer science.

3 Procedure

3.1 Preprocessing

As a data source, the online social media service Twitter is a firehose of information, comprising 16% of all Internet users (Duggan and Brenner, 2013) and millions of "tweets" (140-character posts) per day. Using data from Twitter, Eisenstein et al. (2011) demonstrated an empirical correlation between regional vocabulary and the location of Twitter users. In a similar approach, this project combines metadata of tweets with their content and uses this information to investigate the relationship between AAE syntax and region.

The Twitter data was collected from July to December 2013. We used the website's API that provides a stream of publicly available tweets (approximately 5% of the total tweet volume), restricting our data to geotagged tweets from within the United States. Each tweet includes geographical coordinates (latitude and longitude), name and identity of the Twitter user, and time of creation, as well as its content. The content is broken up and simplified in separate tokens for analysis (e.g. "What's up?" becomes "[what] [' s] [up] [?]"). Following previous work (Eisenstein et al., 2010), we minimize spam posts by removing tweets that contain URLs, and tweets from users that contributed fewer than 20 messages to this data. This gives us a corpus of about 200 million tweets.

Before mining the data, we seek to first eliminate as many retweets as possible to avoid skewing the data. Although we can easily detect retweets that are made through the standard Twitter interface, or are preceded by the token *RT*, we notice that the data contains several unstructured retweets, where a user quotes a tweet from another user without explicitly indicating that it is a retweet. We handle these by simply filtering out every line containing a high-frequency higher order n-gram. After qualitatively observing the results of filtering with different n-gram and frequency combinations, the most efficient and least error-prone filter was determined to be a 6-gram with frequency over 10. Making the assumption that most retweets occur within the same 24-hour period, the tweets of each day were segmented into 6-grams. The 6-grams were tabulated, and all tweets containing a 6-gram with frequency over 10 were omitted. Each day's filtered tweets were then recombined to form the full monthly data. This reduced the size of the corpus by about 26%.

After being filtered, the content of each tweet is fed into a part-of-speech (POS) tagging program developed by Gimpel et al. (2011). This program has achieved over 90% accuracy by using statistics gathered from Twitter data hand-labeled with POS tags. The tagging task is accomplished with a conditional random field using features including non-standard orthography, distributional similarity, and phonetic normalization.

The above uses only 25 tags that range from simple lexemes like O (non-possessive pronoun) to complex morphemes like M (proper noun + verbal). In addition to these basic POS tags, the tweets were tagged with a Penn Treebank-style model trained over another hand-labelled data set (Derczynski et al., 2013). This additional tag set is crucial in detecting constructions like 3rd-person singular -s drop (e.g. "she keep her face down"), which depends on verbal morphology that can be described with PTB tags, but not the simplified tagset of Gimpel et al. (2011).

Owoputi et al. (2013) address the possibility that some AAE tense-aspect-mood (TAM) particles may fall outside the standard POS-tag systems. However, we have observed that "nonstandard" morphemes like *finna* were tagged similarly to Standard American English morphemes, which is likely due to the AAE morphemes exhibiting similar distributional properties to corresponding standard morphemes.

3.2 Querying and Analysis

Using the preprocessed data, it is possible to search through the tagged tweets for a particular syntactic construction by combining the lexical and POS information in a search phrase. For instance, one might use the phrase PRO-ADJ ("we cool," "he cute") to detect copula deletion or PRO-be-V for habitual be. Using regular expressions, these searches can be fine-tuned to ignore noise in the data by searching for patterns like !V-PRO-ADJ ("non-verb+pronoun+adjective"), which ignore false positives like "made me hot." In addition, cases of long-distance constructions like negative concord ("there ain't nobody") can be handled by

Table 1: AAE Constructions and Patterns of Detection

Construction	Example from Corpus	Simplified Pattern	Tagger Used
copula deletion	we stronger than ever	not(V)+PRO+ADJ	PTB
habitual *be*	now i be sober af	not(V)+PRO+*be*+ADJ	PTB
continuative *steady*	steady getting bigger	*steady*+not(N)	Gimpel
completive *done*	u done pissed me off	*done*+V_{PST}	PTB
future *finna (fixing to)*	i'm finna tweet	*finna*+V	Gimpel
remote past *been*	i been had it	PRO/N+*been*+V_{PST}	PTB
negative concord	don't say nothing	*don't/ain't/can't*+V+ *nobody/nothing/nowhere/no*	Gimpel
null genitive marking	time of they life	PRO_{NOM}+N	Gimpel
ass camouflage construction (Collins et al. 2008)	divorced his ass	V+PRO_{POSS}+*ass*	PTB

accounting for a wider context than the keywords themselves, using gaps in the expression. For instance, we detected copula deletion with !V-PRO-ADJ as well as !V-PRO-ADV-ADJ. This strategy was especially useful in preventing false negatives that would otherwise be filtered by rigid patterns (e.g. "he too cute" ignored by !V-PRO-ADJ).

Table 1 contains a list of all constructions queried for this project. To the extent of our knowledge, this is the first study to use regular expressions to use regular expressions and POS tagged data to capture "non-standard" English syntax. The "Tagger" column refers to the POS tagger used to detect the construction: either "Gimpel" (Gimpel et al., 2011) or "PTB" (Derczynski et al., 2013).

Some of the constructions, such as the null genitive (e.g. "time of they life"), could be classified as morphological rather than syntactic phenomena and thus may appear to fall outside the scope of this project. However, it must be noted that these phenomena would not be easily detectable without a POS tagger, which relies on the syntactic context to accurately tag such ambiguous words as "they" (which could be a misspelling of "their"). Furthermore, studies such as Wolfram (2004) that survey AAE grammar also consider morphological phenomena to have comparable frequency and distributional tendencies as syntactic phenomena. Thus, this project chooses to analyze such morphological patterns in the same manner as syntactic patterns.

After querying the data using the regular expressions, the resulting tweets are associated with the metadata corresponding to each tweet. This includes demographic information about the ZIP Code Tabulation Area (ZCTA) associated with the tweet (based on the latitude and longitude coordinates) as well as the estimated gender of the tweeter. ZCTAs are regions defined by the Census Bureau that roughly correspond to postal ZIP codes. Each ZCTA's demographic data includes a number of features. We focus on ethnicity population percentages, overall population in the ZCTA, median age, and percentage of the population living in rented housing (which in some cases could be used to approximate a ZIP code's relative "urban-ness"). The gender of a user is guessed by comparing the tweeter's name with the Social Security Administration's list of baby names from 1995 (`http://www.ssa.gov/oact/babynames/limits.html`), with any user whose name does not appear in the list being assigned a gender of "Unknown". This is a common method used to determine gender in large-scale datasets (Sloan et al., 2013) and one suited to Twitter's younger user base (Duggan and Brenner, 2013).

4 Results

4.1 Comparison of Average Demographics

Our initial approach to the hypothesis – namely, that Twitter shows a quantifiable correlation between ethnicity and usage of AAE syntax – was a comparison of the demographics of the tweeters that use the AAE constructions listed in Table 1 to the average demographics over all users in our data. The constructions' average demographics

Table 2: Mean Demographic Profiles of AAE Construction Users

Construction	User %	Mean % African-American Population	Mean % Caucasian Population	Gender Ratio Female : Male : Unknown
Overall Statistics	1,135,019 users total	$13.67 \pm 18.66\%$	$71.81 \pm 21.66\%$	36.78 : 31.17 : 32.05
Copula Deletion	45.62%	13.64%	71.80%	37.27 : 30.18 : 32.55
ass Camouflage Construction	40.25%	14.09%	71.4%	36.27 : 28.88 : 34.84
Future *finna*	17.33%	14.46%	70.97%	35.37 : 27.65 : 36.98
Habitual *be*	31.63%	14.43%	71.24%	36.04 : 28.44 : 35.52
Continuative *steady*	1.304%	15.45%	69.44%	33.20 : 26.32 : 40.48
Completive *done*	6.061%	14.81%	70.44%	34.06 : 26.95 : 38.98
Remote Past *been*	8.384%	14.83%	70.48%	33.58 : 25.99 : 36.80
Negative Concord	18.14%	14.47%	70.92%	35.30 : 27.70 : 37.00
Negative Inversion	17.66%	14.50%	70.92%	35.30 : 27.63 : 37.07
Null Genitive	13.59%	14.61%	70.75%	34.84 : 27.56 : 37.60

were calculated counting each construction-user only once, regardless of how many times they use that construction.

While reasonable, this approach did not provide encouraging results, as demonstrated by Table 2. The constructions' demographics deviated only slightly from the overall demographics, though the variation reflected the expected trend of higher African-American population (avg. +0.859%) and lower Caucasian population (avg. -0.974%). The constructions showed similar standard deviations to those of the overall demographics. Further ethnic statistics such as average Asian population, which might have been interesting in light of research on dialect reappropriation (Reyes, 2005), were also highly uniform when comparing constructions to overall data.

In addition to ethnic demographics, the gender breakdown was somewhat uninformative as both female and male users were less represented than expected. This may have indicated a failure on the part of the gender-guesser to guess more unusual names like "Notorious J. $tash" that could be associated with AAE syntax. With such negligible deviations from the mean demographics, additional data analysis techniques such as linear regression and clustering of users with similar demographic data would seem to yield negligible results. Thus, these techniques were deemed unnecessary for these averages.

There are a few possible explanations for the inconclusive results in ethnic demographics and gender. First, the information associated with the ZCTA is drawn from the 2010 U.S. census data, which may not match the demographics of social media users. While the time difference between 2010 and 2013 is unlikely to make a significant difference, the discrepancy between real-life statistics and social media metadata may result in statistics contradictory to the Twitter user demographics proposed by Duggan and Brenner (2013). The current study accepts this as a possible source of error and looks toward future studies that directly associate social media users with geographic demographics. More importantly, this thesis relies on ethnic demographics derived from users' environments rather than directly available data such as names, as in Eisenstein et al. (2011). This distinction is crucial, as it dampens the apparent presence of black Twitter users in ZCTAs with low African American population percentages.

While statistically inconclusive for individual constructions, the apparent pervasiveness of AAE syntax as a whole is surprising, even considering the observation by Duggan and Brenner (2013) that 26% of African-American Internet users are on Twitter. Admittedly, no regular expression is free from error, but the apparent 45.62% copula deletion usage rate is impressive for a construction that was once used to parody the speech of AAE speakers (Green, 2002). Furthermore, the users of each construction tend to be located in the data's most common ZCTAs, which are often youth-centric college towns such as San Marcos, Texas. The non-trivial user percentages and significant diffusion of usage outside of expected urban areas build on claims by Wolfram (2004) about "new and intensifying structures in urban AAVE," such as habitual *be*, as well as "receding urban features" such as remote past *been*. The

relatively homogenous distribution of such constructions may even reflect a stable position for AAE as a unified dialect across typical American English dialect regions. However, a long-term Twitter corpus will be necessary to test the diachronic behavior of these apparently "receding" and "intensifying" features.

4.2 Logistic Regression

Following the initial results, we adopted a different approach to measure AAE usage by performing a logistic regression over the demographics collected for the AAE constructions as well as their Standard American English (SAE) counterparts. For example, the SAE equivalent of the AAE future *finna* was considered to be regular genitive pronouns (e.g. AAE "they house" vs. SAE "their house"). At the time of submission, we only extracted SAE demographics for a subset of the constructions. The most salient results of the regression are displayed in Table 3. The variables under consideration are the correlation coefficients relating each construction to the demographics associated with the users, with positive values indicating a trend toward the AAE construction and negative values indicating a trend toward the SAE construction.

Before observing the coefficients, the first notable characteristic of the SAE data is the high rate of occurrence for most standard constructions, such as "Standard *be*+V_{ing}". This may indicate that there is overlap in SAE and AAE usage among Twitter users, which is unsurprising given the prevalence of code-switching among AAE speakers in non-virtual environments (Labov, 2012) as well as the strong potential for dialect spread (Reyes, 2005). To investigate this possibility, future refinement of this regression approximation will compare Twitter users who only employ SAE constructions versus those who only employ the corresponding AAE construction. Though perhaps an artificial distinction that will tend more toward data sparsity than abundance, this strategy will hopefully reveal a split between speakers that tend more toward one dialect than the other, from which further proposals can be tested (e.g. the most reliable construction characterizing each dialect).

The correlation coefficients in Table 3 generally tend toward positive for population of the ZCTA, suggesting a prevalence of AAE in high-population areas and a diffusion of SAE throughout all populated areas. However, the correlation coefficients for Caucasian population and African-American population are less informative and tend slightly toward SAE constructions, with the notable exceptions of negative concord and inversion, which Wolfram (2004) classified as "stable" urban AAE features.

In all cases, the numeric values of the demographic correlation coefficients (including those not shown such as Asian-American population) are so low as to be statistically inconclusive. However, in all AAE/SAE syntax pairs except for the negations, the correlation coefficients for female users showed a tendency toward positive. This could provide support for the female identity-expression hypothesis proposed by Eckert and McConnell-Ginet (2013) but could also indicate an error with the samples obtained using the current AAE syntax patterns (e.g. smaller samples tend to skew toward areas with more women). Further comparison of male vs. female AAE usage is necessary to provide more evidence for the apparent tendency toward women.

5 Conclusion and Future Directions

This thesis proposes (a) a method for detecting AAE syntactic constructions in tweets, and (b) using the metadata from said tweets to approximate the demographics of the users of AAE constructions. The goal of this thesis is to estimate the current state of AAE usage among American social media users. This project has not yet uncovered a clear connection between ethnic demographics and the use of AAE syntax, suggesting that the dialect is more widespread than previous studies such as Wood and Zanuttini (2014) may have predicted. However, several analyses of the data have suggested that women on Twitter employ AAE syntax more than men, even taking into consideration the slightly higher proportion of women using social media. A different approach to data analysis, and potentially stricter syntax-detection patterns (e.g. only detecting special sub-cases of copula deletion), will be necessary to discover trends of AAE usage within the massive dataset.

Since the synchronic approach seemed to yield limited results, the next step in the project will be analyzing the data on a diachronic scale. The first goal of this approach is to corroborate or challenge the claims of Wolfram (2004) concerning

Table 3: Regression Results over AAE and SAE Demographics

AAE/SAE Syntax Pair	SAE User %	Coefficient (Population)	Coefficient (%Caucasian)	Coefficient (%African-American)	Coefficient (Female)
Copula Deletion/ Standard Copula	93.30%	0.0208	−0.0001	−0.0005	0.0321
Future *finna*/ Future *gonna*	61.75%	0.0312	−0.0024	−0.0006	0.0458
Habitual *be*/ Standard *be*+V_{ing}	79.79%	0.0361	−0.0032	−0.0019	0.0529
Continuative *steady*/ Standard *be*+V_{ing}	79.79%	0.0669	−0.0077	−0.0027	0.0505
Completive *done*/ Standard V_{PST}	94.12%	0.0846	−0.0076	−0.0045	0.0685
Negative Concord/ Standard Negation	22.15%	0.0091	0.0009	0.0014	−0.0006
Negative Inversion/ Non-Inverted Negation	20.16%	−0.0181	0.0005	0.0006	0.0018

"intensifying," "stable," and "receding" AAE syntax features by extrapolating a larger pattern of change from the limited time series available (July - December 2013). Secondarily, assuming that some of these features are changing in usage over time, this approach will test whether female Twitter users are leaders of change-in-progress, a trend proven by previous sociolinguistic studies (Eckert and McConnell-Ginet, 2013). In contrast, Reyes (2005) proposes that Asian-American young men adopt AAE slang to emulate African American "hyper-masculinity", a trend which could lead to men rather than women being leaders of dialect reappropriation. To discover such trends of adoption among individual users, it may also make sense to track each tweeter's AAE vs. SAE usage to determine the extent to which an individual user's syntax can change over time.

Outside the scope of this study, future work might consider using a semi-supervised training method over POS n-grams to automatically detect certain syntactic constructions. This would eliminate the need for rigid regular expressions in searching for tweets with AAE syntax, and also enable the detection of a variety of other constructions. In addition, future AAE studies in Twitter may benefit from the approach of Bergsma et al. (2013), which use user names and patterns of interaction to infer "hidden properties" such as gender and race. Under this framework, researchers might leverage online social media metadata to explore emergent linguistic behavior of various speech communities linked by patterns of interaction. This is an intriguing possibility to consider with the increasing presence of online communities like "Black Twitter" (Sharma, 2013), which allow real-world linguistic trends like AAE syntax to propagate in virtual space.

Acknowledgments

This study was funded in part by the Neukom Institute for Computational Science (`http://neukom.dartmouth.edu/`) and the Presidential Scholars Program at Dartmouth. The project is being supervised by Professor James Stanford and postdoctorate scholar Dr. Sravana Reddy of Dartmouth College.

References

Shane Bergsma, Mark Dredze, Benjamin van Durme, Theresa Wilson, and David Yarowsky. 2013. Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter. *Proceedings of NAACL-HLT 2013*, pages 1010–1019.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. *Proceedings of Recent Advances in Natural Language Processing*, pages 198–206.

Maeve Duggan and Joanna Brenner. 2013. The Demographics of Social Media Users - 2012. *Pew Re-*

search Center's Internet & American Life Project, pages 1–14.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and Gender*. Cambridge University Press, New York, 2 edition.

Jacob Eisenstein, Brendan O'Connor, Noah Smith, and Eric Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering Sociolinguistic Associations with Structured Sparsity. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1(49):1365–1374.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. *Proceedings of NAACL-HLT*, pages 380–390.

Lisa Green and Thomas Roeper. 2007. The Acquisition Path for Tense-Aspect: Remote Past and Habitual in Child African American English. *Language Acquisition*, 14(3):269–313.

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

Tyler Kendall, Joan Bresnan, and Gerard van Herk. 2011. The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistics Theory*, 7(2):229–244.

William Labov, Sharon Ash, and Charles Boberg. 2005. *The Atlas of North American English: Phonetics, Phonology, and Sound Change*. Walter de Gruyer, Berlin.

Wiliam Labov. 2012. *Language in the inner city: Studies in the black English vernacular*. University of Philadelphia Press, Philadelphia, PA.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. *Proceedings of NAACL-HLT 2013*, pages 380–390.

Angela Reyes. 2005. Appropriation of African American slang by Asian American youth. *Journal of Sociolinguistics*, 9(4):509–532.

Sanjay Sharma. 2013. Black Twitter?: Racial Hashtags, Networks and Contagion. *new formations: a journal of culture/theory/politics*, 78(1):46–64.

Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. 2013. Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online*, 18(3).

Walt Wolfram. 2004. Urban African American Vernacular English: morphology and syntax*. In Bernard Kortmann, editor, *A handbook of varieties of English. 1. Phonology, Volume 2*, volume 2, pages 319–340. Walter de Gruyer.

Jim Wood and Natalie Zanuttini. 2014. The Yale Grammatical Diversity Project. http://microsyntax.sites.yale.edu/.

Malcah Yaeger-Dror and Erik R. Thomas. 2010. *African American English Speakers and Their Participation in Local Sound Changes: A Comparative Study*. Duke University Press for the American Dialect Society, Durham, NC.