Extracting Non-Standard Varieties via the Twitter API: The Case of AAE

Kimberley Baxter

The present paper examines methodology in the use of Twitter in the corpus-based analysis of African American English (AAE) syntax. 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. Twitter's API allows access to their full database of tweets, which is a much larger and more accessible dataset than its large social media contemporaries.

There are well-documented challenges in analyzing non-standard varieties of English commonly used in social media (Plank, 2016), and AAE is no different. Efforts to normalize non-standard varieties often do not allow for the extraction or analysis of non-standard language as it occurs among speakers. In the case of AAE, standard methods of extracting data via Twitter's API are insufficient, lacking the specifications necessary to isolate certain parts of speech exclusive to AAE, and differentiate them from similar lexical items in Mainstream American English (MAE). Despite habitual be and copula/auxiliary be having two separate uses, the word "be" and its conjugations look and sound exactly the same, as seen below in examples (1) and (2).

(1) He be running. (Meaning: He tends to run. He usually runs. He is not necessarily running right now.)

(2) He is running. (He is currently running right now.)

This results in a high number of false positives and data which is rendered unusable due to the sheer size of the dataset, which may number in the millions of tweets, thus rendering the manual elimination of false positives unfeasible. This project ultimately aims to produce an alternative, syntax-based method which allows the user to eliminate a great deal of the aforementioned false positives by coding the syntactic constraints of this part of speech, thus allowing for the extraction of non-standard varieties, which would otherwise be inaccessible due to the comparative lack of specialized part of speech taggers designed for this task. While the initial focus is on Twitter, this tool will ultimately combine a range of methodological approaches and hopes to foster collaboration between researchers.