Analysis of billions of Twitter words reveals how American English develops

September 26 2018

Credit: CC0 Public Domain

Linguists and geographers analysed 8.9 billion words contained within 980 million Tweets posted across the United States between 2013 and 2014 to identify the regions from which new words tend to originate.

Led by Professor Jack Grieve, from the Centre for Corpus Research at the University of Birmingham, researchers used advanced computer technology to analyse the geocoded Tweets which revealed the precise longitude and latitude of the user at the time of posting.

They tracked the origin of 54 newly emerging words in American English. For example, they found that the word 'baeless', which mean 'to be single', originated from Deep South, while the word 'mutuals', which is short for 'mutual friends', originated from the West Coast.

Geo-coded data from Twitter allowed them to create maps for these 54 words, showing how the phrases had spread across the country over time.

Applying modern computational techniques to the study of language variation and change, the team identified that development of new words in Modern American English centred on five regions: The West Coast, the Northeast, the Mid Atlantic, the Deep South, and the Gulf Coast.

Professor Grieve commented: "This is the first time that such a large sample of emerging words or any type of linguistic innovation has been mapped in one language. Twitter is only one variety of language, but given that almost all these words are used in everyday speech, we believe our results reflect the words' general spread in American English.

"Our study provides a framework for future research by showing how the origin and spread of emerging words can be measured and mapped. Linguistics is shifting from a social science to a data science, where linguists are increasingly analysing massive amounts of natural language harvested online.

"This is allowing us to pursue new research questions that would have been impossible to investigate just a few years ago. We can analyse in very fine detail how language changes over short periods of time and understand the processes through which languages evolve—one of the most challenging questions in science."

The researchers' findings also challenge existing theories of the spread of new words. They show that new words do not simply spread out unconstrained from their source, nor do they spread from one large city to the next, as predicted by previously developed theories for the spread of new words, known as the 'wave' and 'gravity' models.

Instead, the study found the spread of new words is constrained by cultural patterns. New words tend to spread within cultural regions, before reaching the rest of the United States. It also found that African American English was a major source of lexical innovation on US Twitter.

Professor Grieve is speaking about the team's research at the New Ways of Analyzing Variation (NWAV) conference held at New York University from October 18 to 21. He will focus on how these words spread just in New York City over the time period in question, as well as delivering a workshop on 'computational sociolinguistics'.

More information: Jack Grieve et al, Mapping Lexical Innovation on American Social Media, *Journal of English Linguistics* (2018). DOI: 10.1177/0075424218793191

Provided by University of Birmingham