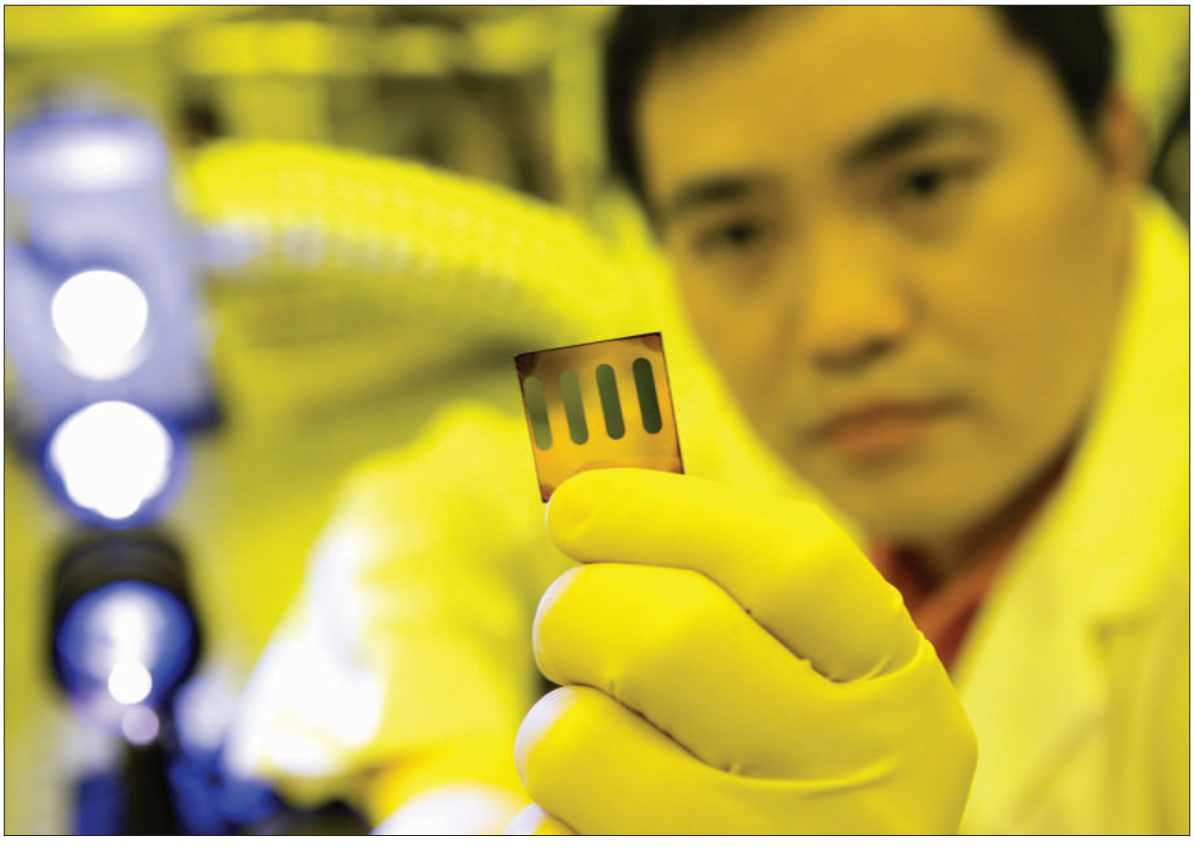

Singapore Journal of

Scientific Research

ISSN: 2010-006x

science alert

http://scialert.net/sjsr

Regional Dialects Are Alive and Well on Twitter: Slang Terms Like Y'all, Yinz, Koo, Coo and Suttin Predict Location of Tweet Authors

Microbloggers may think they're interacting in one big Twitterverse, but researchers at Carnegie Mellon University's School of Computer Science find that regional slang and dialects are as evident in tweets as they are in everyday conversations.

Postings on Twitter reflect some well-known regionalisms, such as Southerners' "y'all," and Pittsburghers' "yinz," and the usual regional divides in references to soda, pop and Coke. But Jacob Eisenstein, a Post-doctoral Fellow in CMU's Machine Learning Department, said the automated method he and his colleagues have developed for analyzing Twitter word use shows that regional dialects appear to be evolving within social media.

In northern California, something that's cool is "koo" in tweets, while in southern California, it's "coo." In many cities, something is "sumthin," but tweets in New York City favor "suttin." While many of us might complain in tweets of being "very" tired, people in northern California tend to be "hella" tired, New Yorkers "deadass" tired and Angelenos are simply tired "af."

The "af" is an acronym that, like many others on Twitter, stands for a vulgarity. LOL is a commonly used acronym for "laughing out loud," but Twitterers in Washington, D.C., seem to have an affinity for the cruder LLS.

Eisenstein said some of this usage clearly is shaped by the 140-character limit of Twitter messages, but geography's influence also is apparent. The statistical model the CMU team used to recognize regional variation in word use and topics could predict the location of a microblogger in the continental United States with a median error of about 300 miles.

Eisenstein is presenting the study on Jan. 8 at the Linguistic Society of America annual meeting in Pittsburgh. The paper is available online at http://people.csail.mit.edu/jacobe/papers/emnlp2010.pdf.

Studies of regional dialects traditionally have been based primarily on oral interviews, Eisenstein said, noting that written communication often is less reflective of regional influences because writing, even in blogs, tends to be formal and thus homogenized. But Twitter offers a new way of studying regional lexicon, he explained, because tweets are informal and conversational. Furthermore, people who tweet using mobile phones have the option of geotagging their messages with GPS coordinates.

For this study, Eisenstein and his co-authors -- Eric P. Xing, Associate Professor of Machine Learning, Noah A. Smith, Assistant Professor in the Language Technologies Institute (LTI), and Brendan O'Connor, machine learning graduate student -- collected a week's worth of Twitter messages in March 2010, and selected geotagged messages from Twitter users who wrote at least 20 messages. That yielded a data base of 9,500 users and 380,000 messages.

Though the researchers could pinpoint the users' locations using the geotags, they can only guess as to their profiles. Eisenstein said it's reasonable to assume that people sending lots of tweets from mobile phones are younger than the average Twitter user and the topics discussed by these users seem to reflect that.

Automated analysis of Twitter message streams offers linguists an opportunity to watch regional dialects evolve in real time. "It will be interesting to see what happens. Will 'suttin' remain a word we see primarily in New York City, or will it spread?" Eisenstein asked.

It might be a mistake to assume that the greater interconnectivity afforded by computer networks and sites such as Twitter will necessarily result in more homogeneity

in language. The social circles maintained by social networks such as Twitter often are geographically focused, he noted. Also, many people use the Internet to seek out like-minded people with similar interests, rather than expose themselves to a broader range of ideas and experiences.

The research was supported, in part, by funding from Google, the Air Force Office of Scientific Research, the Office of Naval Research, the National Science Foundation and the Alfred P. Sloan Foundation.

Source: The above story is reprinted from materials provided by Carnegie Mellon University, via EurekAlert!, a service of AAAS.