Researchers Discover "Superdialects" Through Twitter Study

Using Twitter to learn how people speak differently.
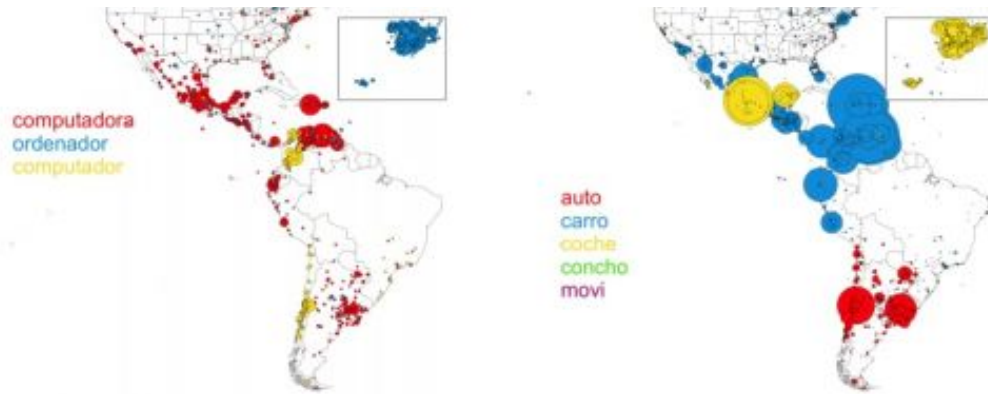
SELENA LARSON · AUG 7, 2014

A team of European researchers have discovered distinct differences in the dialects of the Spanish language by studying it through 140-character bursts.

In the study Crowdsourcing Dialect Characterization through Twitter, Bruno Gonçalves from Toulon University in France and David Sánchez from the Institute for Cross-Disciplinary Physics and Complex Systems in Spain, studied 50 million geo-located Spanish tweets over the course of three years. They discovered that, as expected, most came from Spain, Spanish America, and across the U.S., as first reported by the MIT Technology Review.

To take a closer look at how words and language vary across geographic locations, the team studied how Spanish words for the same object can vary based on where people tweet from—for instance, the word for "car" can be *auto*, *carro*, *coche*, *concho*, or *movi*. The researchers discovered that some expressions are clustered in different regions, demonstrating different geographic dialects.

computadora
ordenador
computador

auto
carro
coche
concho
movi

"Up until now researchers had been limited to small scale surveys and in-person interviews with the obvious limitations in terms of number of individuals considered and the associated costs of travel," Bruno Gonçalves said in an email interview with ReadWrite. "With Twitter and the advent of cheap GPS-enabled smartphones, we can study how language is used by millions of users scattered across the world in their day to day communications."

Thanks to the huge quantity of data, Gonçalves and Sánchez discovered that the Spanish language is split between two "superdialects," a form of language that includes two or more dialects. One Spanish superdialect that's spoken in large American and Spanish cities, and one that's common among rural parts of Spanish-speaking countries.

"It was relatively well known that some expressions were more localized than others, and we were hoping to be able to see that in the data," Gonçalves said. "The urban/rural difference had not been observed before as traditionally dialect researchers have focused less on city dwellers."

While this study is the first to discover such superdialects by analyzing tweets, it's not the first time linguists have turned to the social network to figure out how people talk.

Scientists have used Twitter to figure out how slang words and phrases like "bro" and "AF" originated in different geographic areas, and to study the evolution of languages, like Welsh, in real-time.

"This is also a good example of the potential benefits of big data for the study of human behavior," Gonçalves said. "Data that was generated by users when they used the Twitter platform to communicate with their followers and friends allowed us to study how a specific language is used 'in the wild'."

Through similar research, scientists could discover "superdialects" in other languages.

Lead image by Andreas Lehner. Map courtesy of the study.

Read more from Selena Larson.

Aug 7, 2014

/ Language study / Superdialects / Twitter