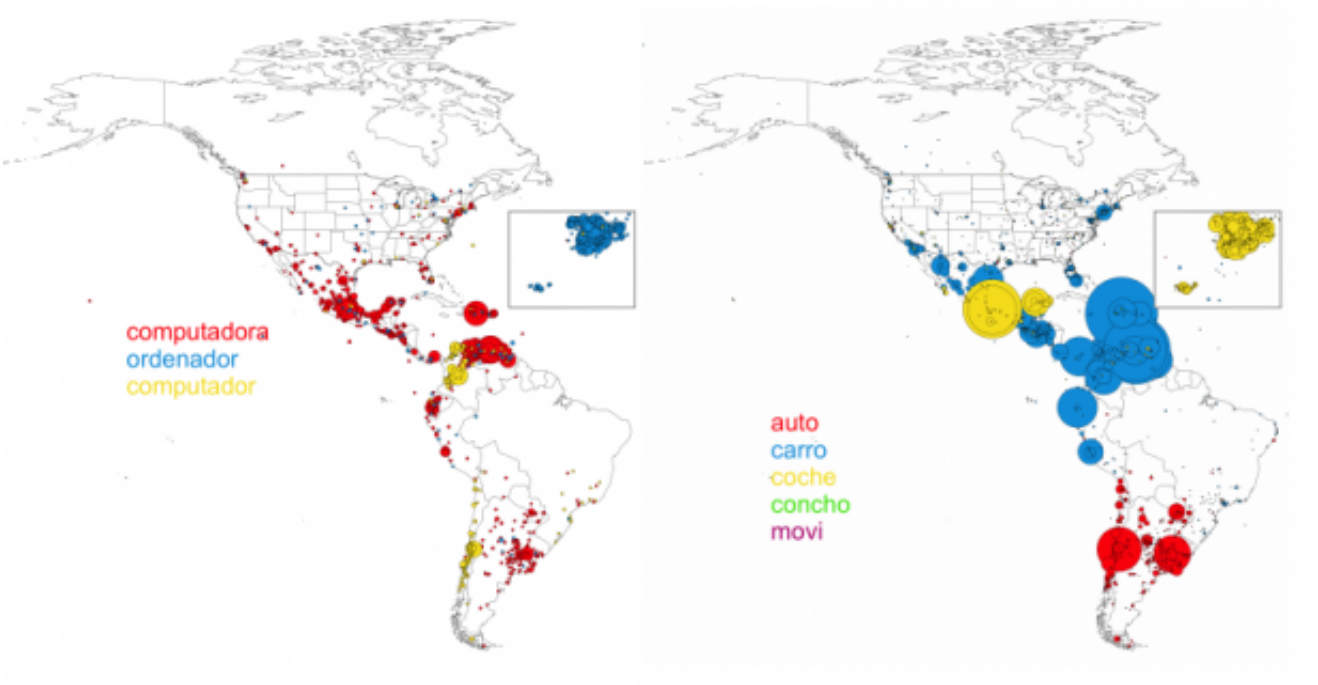
MIT Technology Review

Emerging Technology From the arXiv
August 7, 2014

Computational Linguistics of Twitter Reveals the Existence of Global Superdialects

The first study of dialects on Twitter reveals global patterns that have never been observed before.



A dialect is a particular form of language limited to a specific region or social group. Linguists are fascinated by dialects because they reveal social classes, patterns of immigration and how groups have influenced each other in the past.

But studying dialects is hard work. Traditionally, linguists do this by interviewing a relatively small number of people, typically a few hundred, and asking them to fill out questionnaires. Researchers then use the results to create linguistic atlases but these are naturally limited by the choice of the locations and individuals who have been studied.

Today, Bruno Gonçalves at the University of Toulon in France and David Sánchez at the Institute for Cross-Disciplinary Physics and Complex Systems on the island of Majorca, Spain, say they have found a new way to study dialects on a global scale using messages posted on Twitter. The results reveal a

major surprise about the way dialects are distributed around the world and provide a fascinating snapshot of how they are evolving under various new pressures, such as global communication mechanisms like Twitter.

Gonçalves and Sánchez begin by sampling all of the tweets written in Spanish over two years and that also contain geolocation information. That gave them a database of 50 million geolocated tweets, with most from Spain, Spanish America, and the United States.

They then searched these tweets for word variations that are indicative of specific dialects. For example, the word for car in Spanish can be auto, automóvil, carro, coche, concho, or movi, with each being more common in different dialects. Different words for bra include ajustador, ajustadores, brasiel, brassiere, corpiño, portaseno, sostén, soutien, sutién, sujetador, and tallador while variations on computer include computador, computadora, microcomputador, microcomputadora, ordenador, PC, and so on.

They then plotted where in the world these different words were being used, producing a map of their distribution. This map clearly shows how different words are commonly used in certain parts of the world.

However, they also looked at the environments in which the words were used, whether in large cities or in rural locations. And that revealed a major surprise.

It turns out that Spanish dialects falls into two major groups which Gonçalves and Sánchez call superdialects. The first of these is used more or less exclusively in major Spanish and American cities. This is an international variety of Spanish that is similar across continents. Gonçalves and Sánchez speculate that this is the result of an increasing homogenization of language caused by global communication systems like Twitter.

The second superdialect is used almost exclusively in rural areas. Gonçalves and Sánchez used a machine learning algorithm to find subclusters within this group and discovered three different variations. These correspond to a dialect used in Spain, a Caribbean and Latin American dialect and another variation used exclusively in South America.

The researchers say these regions reflect the settlement patterns of Spanish immigrants dating back many centuries. “Conquerors and settlers occupied first the territories of Mexico, Peru and the Caribbean, and only much later colonists established permanent residence in [South America], which stayed away from prestigious linguistic norms,” they say.

The fact that patterns of language have preserved this history is fascinating. “This strong cultural heritage that can still be observed, centuries later, in our datasets deserves to be further analyzed in future works,” say Gonçalves and Sánchez.

That is important work that reveals the existence of superdialects on a global scale for the first time. It also demonstrates the power of computational linguistics and how it can be applied to modern forms of communication such as Twitter to reveal patterns on an unprecedented scale.

There is clearly plenty of low hanging fruit in this area although Gonçalves and Sánchez warn that some languages will still be difficult to study in this way, for example Mandarin because speakers do not have easy access to Twitter.

Nevertheless, expect to see a lot more from these kinds of computational linguistic techniques in the not too distant future.

Ref: arxiv.org/abs/1407.7094 : Crowdsourcing Dialect Characterization through Twitter

Tagged: Communications