

We present a new computational technique to detect and analyze statistically significant geographic variation in language. Our meta-analysis approach captures statistical properties of word usage across geographical regions and uses statistical methods to identify significant changes specific to regions. While previous approaches have primarily focused on lexical variation between regions, our method identifies words that demonstrate semantic and syntactic variation as well. We extend recently developed techniques for neural language models to learn word representations which capture differing semantics across geographical regions. In order to quantify this variation and ensure robust detection of true regional differences, we formulate a null model to determine whether observed changes are statistically significant. Our method is the first such approach to explicitly account for random variation due to chance while detecting regional variation in word meaning. To validate our model, we study and analyze two different massive on-line data sets: millions of tweets from Twitter spanning not only four different countries but also fifty states, as well as millions of phrases contained in the Google Book Ngrams. Our analysis reveals interesting facets of language change at multiple scales of geographic resolution – from neighboring states to distant continents