Analysis of Political Texts in Multiple Languages

Burt L. Monroe

Liberal Arts Professor of Political Science, Social Data Analytics, and Informatics
Director, Center for Social Data Analytics
Penn State College of Liberal Arts

Abstract:

Social scientists confront a fundamental barrier to conducting cross-national text-as-data research: multiple languages. Many text analytic techniques leverage the repeated co-occurrence of words, but most language pairs have very few words in common. Building on recent advances in transfer learning and multilingual word embeddings (and old advances in multivariate statistics) we propose a new method for analyzing multilingual corpora, which does not rely on translation to a common language or a polyglot researcher or team. Further, we propose novel methods for extracting topics and scales from multilingual corpora. We demonstrate the validity and utility of these methods to comparative politics through analyses of a corpus of parliamentary speeches from eight European languages.

Bio:

Burt Monroe is Liberal Arts Professor of Political Science, Social Data Analytics, and Informatics at Penn State. He is founding Head of the Program in Social Data Analytics [SoDA], which offers interdisciplinary undergraduate and doctoral degrees, and founding Director of the Center for Social Data Analytics. The SoDA degrees and Center are products of the NSF-funded Big Data Social Science IGERT, which he directed from 2012-2018. He is also Chief Scientist of the McCourtney Institute for Democracy's Mood of the Nation Poll, directing efforts to understand the poll's open-ended survey questions through text analytics.