Title: Coh-MetrixML: Multilingual Semantic and Syntactic Analysis
Authors: Leah Windsor (IIS, UofM), Grayson Cupit (IIS, UofM), Zhiqiang Cai (IIS, UofM), Andrew Tackett (IIS, UofM), and Alistair Windsor (Math &IIS, UofM)

In this paper, we propose a computational solution for deficiencies in the data generating process (DGP): a multilingual syntax and semantics analysis tool called Coh-MetrixML. Discourse analysis, broadly defined, skews heavily toward English-language sources and reflects a Western bias that influences the scope, interpretation, and external validity of research on international politics (1). Most corpora for social and political research are in English or come from Western European languages. On the other hand, data collected in the developing world is often not gathered with academic use in mind. Professional translation services are not only time-consuming, but also expensive. Translation may provide the gist of a document, but this is no substitution for the ability to preserve the structure and content of the original language. Importantly, many computational linguistics programs are often available only for English-language corpora. These issues contribute to selection bias along the entire "value chain" of corpus production that affect research outcomes and the conclusions we draw about socio-political problems in the world.

Recent research has demonstrated how bag-of-words approaches such as sentiment analysis and topic modeling can generate information about the content and categories of socio-political problems from multilingual, and mono-lingual multicultural corpora. Linguistic style is a valuable and reliable indicator of many social and behavioral elements, such as status, hierarchy, cognitive function, personality, persuasive strategy, and deception. Coh-Metrix is a computational linguistics program developed to analyze the syntactic and semantic properties of natural language (2). It measures features of language such as passive voice, syntactic complexity, narrativity, and two types of cohesion - local and global. We extend the multilingual analytic capacity to other languages, including French, Spanish, Chinese, German, and Arabic with the program Coh-MetrixML (multi-lingual). By extending the semantic and syntactic analytic capacity of computational linguistics tools, Coh-MetrixML will make more representative corpora available to researchers.

From a software architecture perspective, Coh-MetrixML is written with the aid of the Cats and FS2 (Functional Streams for Scala) libraries (3). The stream is a fundamental data type within CMXML. Text documents are loaded from the disk as a stream of filepaths. Through the use of pipes, the stream of filepaths is transformed into a stream of files, then a stream of text, then a stream of data-enriched objects with various CoreNLP and WordNet features, then finally through various calculators to a stream of results (4–6). The result stream is consumed by an in-memory SQL database (H2) where results are stored until all the analysis is complete and the output file can be written. This streaming approach represents improvements over earlier versions of Coh-Metrix in a variety of ways. Furthermore, the inherently modular philosophy of streams and pipes makes the codebase easily extensible with additional calculators, pre-processing stages, or custom CoreNLP models.

Works Cited