„Automated Text Analysis and the Study of Constitutions“

David Law
(University of Virginia)

Lawyers know that the choice of language in legal texts matters. This is particularly true of constitutions, which are not simply transmission mechanisms for legal rules and concepts. Constitutional language is also a medium of emphasis, tone, rhetoric, and style; it bears the semantic footprint of its author and the forces that have shaped its authorship. Linguistic patterns are telltale indicators not only of substantive topics, but also different genres or styles of constitutionalism and competing influences on constitution-writing.

Systematic empirical analysis of legal language has long been stymied by the fact it is difficult for human readers to discern subtle or complex linguistic patterns with accuracy and consistency across a large corpus of documents, such as the hundreds of constitutions that have been adopted over the last two centuries. The traditional way in which quantitative empirical scholars analyze legal texts is to code the text into numerical data that can then be analyzed statistically. In the process of coding language into numbers, however, the text itself is necessarily discarded, along with anything that falls outside the coding scheme. Consequently, traditional coding-based approaches are not well suited to studying phenomena that escape our awareness or cannot easily be coded by hand.

Recent innovations in the area of automated content analysis have made it possible to perform precisely this kind of analysis. These new methodologies, adopted from computational linguistics, excel at identifying and analyzing subtle, complex verbal patterns in a rapid, systematic, and objective way. They potentially enable us to measure, in quantitative terms, how much of the text is associated at the linguistic level with a particular topic, influence, or genre. Although automated content analysis has already attracted widespread interest for its obvious advantages of speed and consistency over manual techniques, its most profound advantage may be that it liberates us from reliance on potentially incomplete coding schemes. Instead of forcing us to analyze text through the filter of our preexisting conceptual categories, automated content analysis allows the text to speak for itself.