Over the past few years, natural language processing (NLP) has become an increasingly important element in computational research in the humanities and social sciences, enabling sophisticated analyses that can go far beyond simple word counting. However, there is a substantial gap between the quality of the NLP used by researchers in the humanities and the state of the art. NLP research has overwhelmingly focused not only on one language (English) but also one domain (newswire)---leaving many other languages, dialects and domains (such as literary text) underserved.

In this talk, I’ll advocate for two things that I think are necessary to drive the next generation of textual work in the computational humanities. First, I’ll argue for the importance of structured linguistic representations in computational models of text, surveying several recent projects that have leveraged that structure to good effect. Second, I’ll advocate for the development of high-quality NLP for the long tail of languages, dialects and domains that humanists study--and which humanists are in the best position to take the reins and make progress on. By leveraging standard machine learning techniques with disciplinary expertise only humanists can provide, we can both dramatically expand the scope of NLP to be applied to a much wider variety of texts in our cultural record and use the linguistic structure we infer to help define new tasks altogether.