Jan Rybicki

What Else Can Be Done with Lots of Texts: Distant Reading by Counting Words, Lemmas and Part-of-Speech Tags

Computational linguistics and corpus linguistics use huge and structured sets of language material to study the behaviour of linguistic elements in their natural habitat. This material is annotated and searchable; corpora help understand the nature of language at many levels of abstraction: vocabulary, syntax, grammar. Yet large quantities of linguistic material are also studied in other fields, usually associated with cultural and literary studies. In particular, the combined use of the macroscopic Distant Reading approach to culture with the microscopic focus on frequencies of simple linguistic features such as words, lemmas or parts of speech as practiced in (literary) stylometry may be an exciting complementation to traditional literary studies and, at the same time, provide new insights into the functioning of language. For one, authorship attribution based on such purely linguistic data bears testimony to the strength of evidence derived from large-scale collections of texts. Interestingly, too much information added onto plain text may at times distort the various "signals" discernible through quantitative study: those of authorship, chronology, genre, or theme. This presentation uses a selection of case studies to illustrate these phenomena.