

Brad Pasanek
Postdoctoral Fellow
Annenberg Center for Communication, USC
bpasanek@annenberg.edu

D. Sculley
Graduate Student
Department of Computer Science, Tufts University
dsculley@cs.tufts.edu

Mining Millions of Metaphors

One of the key problems in any research is to choose the appropriate scale of analysis -- are we looking out into the heavens, or down into atoms? To conceive a digital library as a collection of a million books may restrict analysis to only one level of granularity. In this paper we examine the consequences and opportunities resulting from a shift of scale, where the desired unit of interpretation is something smaller than a text: it is a keyword, a motif, or a metaphor. A million books distilled into a billion meaningful components become raw material for a history of language, literature, and thought that has never before been possible. While books herded into genres and organized by period remain irregular, idiosyncratic, and meaningful in only the most shifting and context-dependent ways, keywords or metaphors are lowest common denominators. At the semantic level—at the level of words, evocative images, and metaphors—long term regularity and patterns emerge in collection, analysis, and taxonomy.

This paper follows the foregoing course of thought through three stages: first, the manual curation of a high quality database of metaphors; second, the expansion of this database through automated and human-assisted techniques; finally, experiments and opportunities for the application of machine learning, data mining, and natural language processing techniques to help find patterns and meaning concealed at this important level of granularity.

In the first section, we introduce *The Mind is a Metaphor* database (<http://mind-metaphor.stanford.edu>), which houses a collection of over 7,900 metaphors of mind. The database is an integral part of Brad Pasanek's dissertation, *Eighteenth-Century Metaphors of Mind, A Dictionary* (completed summer 2006). Pasanek's project was made possible by searching the Chadwyck-Healey databases, ECCO, EEBO, and other collections of electronic texts and is among the first dissertations to use electronic texts as a target of literary-historical scholarship. In Chadwyck-Healey alone, evidence was gathered from 877 books of poetry by 247 different poets, 96 long prose narratives, and 628 plays by 137 different playwrights. This process was labor intensive, depending on years of hand-labeling by a human expert, but resulted in an extremely high quality database of curated metaphors that serves as the ground truth for both automated expansion and data mining tasks.

This initial database of eighteenth-century metaphors lays the foundation for a more ambitious research project. In the second section of the paper, we discuss our attempts to automate the collection of several thousand more metaphors in neighboring periods and examine opportunities for interdisciplinary collaboration enabled through online community involvement and automation with human assistance. Our goal is to assemble a vast database of metaphors harvested from Homeric epics, postmodern cyberpunk novels, and everything in between. Our choice of metaphors of *mind* is then a studied one, as some concept of the self, the mind, or the soul is readily located in most every culture and historical period. Cognitive and computational linguists, rhetoricians, literary critics, intellectual historians, psychologists, philosophers, and neuroscientists would all profit from browsing through this long, deep history of metaphors for the mind.

In the third and final section of the paper, we demonstrate the value of examining text on the metaphorical scale by reporting some of our initial findings and analysis on this data, using a variety of statistical and data-mining techniques. For example, the evidence amassed suggests that the eighteenth-century was not a period in which John Locke's use of the Tabula Rasa metaphor banished innate ideas from public discourse; the evidence troubles M. H. Abrams' famous argument that mirror metaphors of mind gave way to lamp metaphors in eighteenth-century literature, and it indicates that in a century marked by volatile, world-historical revolutions only a few, surprising clusters of metaphors exhibit significant change. Indeed, analysis of this collection of thousands of metaphors upends much of what we thought we knew about eighteenth-century intellectual history. Mining a database of a million metaphors, or more, could enable revolutionary insights across a range of fields.