Trimming time in the stacks

December 21 2011, By Nicole Freeling

A sophisticated text-analyzing tool developed by a UC Berkeley graduate student could speed literary searches for humanities scholars and other researchers.

For graduate students in the humanities, spending months or years in a library combing through obscure texts is a time-honored, if not always savored, rite of passage. Technologies like Google Scholar have speeded the process somewhat, but students still must sift through tens of thousands of results to pluck a few kernels of useful research from the dross.

Now, as part of her doctoral thesis in computer science, UC Berkeley graduate student Aditi Muralidharan has developed a technology platform that she believes has the potential to transform the process of scholarly research. It could reduce what takes months in the stacks and

days in front of a computer screen into an electronic query that takes about five minutes.

Called WordSeer, the program uses rubrics about the way we use and structure language to bring a facsimile of human logic to the business of interpreting search results.

“Up to now, the state of art for search in literary scholarship has been to ask a graduate student,” Muralidharan says. That’s because existing technologies lack the intuitive understanding required for meaningful analysis of text.

“Search technology works much better when the target is a known object or piece of data,” she said. “But humanities scholars don’t know what the item they’re looking for is exactly. It could be a document, a paragraph or a sentence. The target is much more obscure.“

WordSeer employs a technology called Natural Language Processing, which uses understanding about parts of speech, usage and word relationships to enable search that is both more broad — in that it allows users to cast about for citations without a key word — and more specific, in that it makes determinations about the relevancy of text.

The technology itself is about 20 years old but, ironically, the programming language it is based on has itself not been translated into plain English. Thus far it has been executable only by use of a complex programming code.

Muralidharan has been able to create a simple user interface for the technology, allowing users to pose questions of particular sets of text. Berkeley English professors Bryan Wagner and Todd Carmody, for example, employed WordSeer to assist with research exploring American slave narratives to better understand slaves’ relationships with

God.

They posed two questions of the collection (which had been digitized): “What does God do?” and “How was God described?” The query returned thousands of citations and ranked the results by frequency of usage. The search revealed that positive attributes — such as great, wise and merciful — and benevolent actions — such as bless, give and grant — were in the overwhelming majority.

“You don’t have to read all 4,000 matches to get a sense of the overall tone of the collection,” said Muralidharan. In this case, while one might expect slaves in their misery to blame God, the search results indicated the relationship was quite a positive one.

“The WordSeer project is one of the most sophisticated tools for computational text analysis I've seen,” said Wagner. “There are many tools that tabulate words and phrases, but WordSeer can discern grammatical structure as well as stylistic features, representing the results in truly interactive visualizations.”

Funded through a grant from the National Endowment for the Humanities, Muralidharan believes the technology could be useful for all kinds of research that depends upon poring over vast quantities of text, including journalistic and legal research. Berkeley School of Information professor Marti Hearst, Muralidharan’s faculty adviser on the project, said this kind of technology has the potential not just to speed literary research, but to influence how scholars make sense of their subject matter.

“Humanities scholars will continue to formulate their hypotheses as they have before, but will also get new ideas by being inspired by patterns in the text that these tools suggest,” Hearst said. “These tools will also help scholars test their hypotheses across many more documents than they can

do now manually.”

As a condition of the grant, the program will be freely available and open source. Muralidharan says her intention is not to sell the technology. Rather, it’s to complete her doctoral thesis — and along the way, smooth the process of doing research for some of her fellow scholars.

Provided by University of California - Berkeley