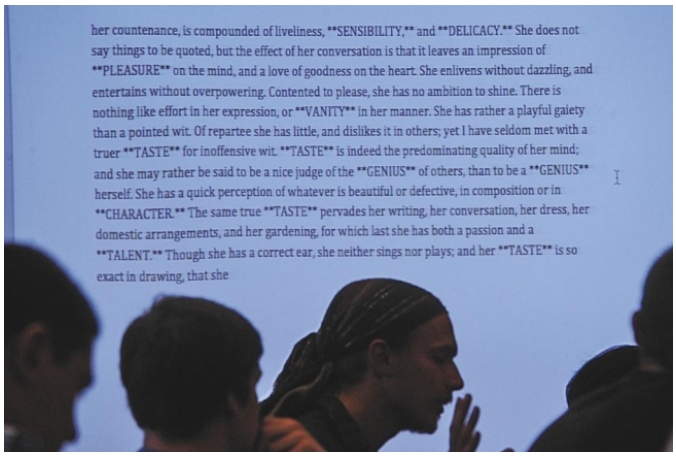
THE CHRONICLE OF HIGHER EDUCATION

TECHNOLOGY

The Humanities Go Google



Noah Berger for The Chronicle

Students from Stanford U.'s Literature Lab analyze text from an 1809 novel by Hannah More.

By Marc Parry | MAY 28, 2010

PALO ALTO, CALIF.

Matthew L. Jockers may be the first English professor to assign 1,200 novels in one class.

Lucky for the students, they don't have to read them.

As grunts in Stanford University's new Literature Lab, these students investigate the evolution of literary style by teaming up like biologists and using computer programs to "read" an entire library.

It's a controversial vision for changing a field still steeped in individual readers' careful analyses of texts. And it could become a more common way of doing business in the humanities as millions of books are made machine-readable through new tools like Google's digital library. History, literature, language studies: For any discipline where research focuses on books, some experts say, academe is at a computational crossroads.

Data-diggers are gunning to debunk old claims based on "anecdotal" evidence and answer once-impossible questions about the evolution of ideas, language, and culture. Critics, meanwhile, worry that these stat-happy quants take the

human out of the humanities. Novels aren't commodities like bags of flour, they warn. Cranking words from deeply specific texts like grist through a mill is a recipe for lousy research, they say—and a potential disaster for the profession.

The debate over the value of the work at Stanford previews the disciplinary battles that may erupt elsewhere as Big Data bumps into entrenched traditions. It also underscores complications colleges encounter when they pin their digital dreams on a corporation.

Authors and publishers have besieged Google's plan to digitize the world's books, accusing the company of copyright infringement. The legal limbo that has tied up a settlement of their lawsuits is hanging a question mark over universities' plans to build centers for research on the books Google scanned from their libraries. Another complication: Worrisome questions remain about the quality of Google's data, which may be less like the library of Alexandria and more like a haphazardly organized used-book shop.

But at Stanford, legal and technical headaches may be worth the sweeping rewards of becoming one of perhaps two places in the world to host the greatest digital library ever built. The university is planning to chase that prize—and the prestige, recruitment power, and seminal research that could come with it. So is HathiTrust, a digital library consortium whose leaders include the University of Michigan at Ann Arbor, Indiana University at Bloomington, and the University of California system.

"It's like the invention of the telescope," Franco Moretti, a Stanford professor of English and comparative literature, says of Google Books. "All of a sudden, an enormous amount of matter becomes visible."

Once scholars like Mr. Moretti can gaze into those new galaxies, however, they'll have to answer the biggest question of all: So what?

Partners in Provocation

"Lab" is a generous description for the place where Mr. Moretti and his disciples work on these problems. The research effort he leads with Mr. Jockers, a lecturer and academic-technology specialist in the English department, is housed in an ugly room the size of one professor's office. The lab has no window and nothing on the walls but some whiteboards scrawled with algorithms.

Like others itching to peer into Google's unfinished telescope, Mr. Moretti and his colleagues here are honing their methods with home-grown prototypes. One lesson they've learned is you can't do this humanities research the old way: like a monk, alone.

You need a team. To sort, interrogate, and interpret roughly 1,000 digital texts, scholars have brought together a data-mining gang drawn from the departments of English, history, and computer science. They're the rare clique of humanities graduate students who work across disciplines and discuss programming languages over beer, an unlikely mix of "techies" and "fuzzies" with enough characters for a reality-TV show.

Their backbone is Mr. Jockers, an obsessive tech whiz from Montana who has run a 50-mile race in his spare time and gets so excited talking about text-mining that his knees bob up and down. Mr. Moretti is his more conservative partner in provocation, a vest-and-spectacles-wearing Italian native who tempers Mr. Jockers's excitement with questions and punctuates sentences with his large hands. In their role as Lewis and Clark on the literary frontier, the duo have a penchant for firing shots at the establishment; Mr. Moretti once told *The New York Times* that their field is in some ways one of "the most backward disciplines in the academy."

The idea that animates his vision for pushing the field forward is "distant reading." Mr. Moretti and Mr. Jockers say scholars should step back from scrutinizing individual texts to probe whole systems by counting, mapping, and graphing novels.

And not just famous ones. New insights can be gleaned by shining a spotlight into the "cellars of culture" beneath the small portion of works that are typically studied, Mr. Moretti believes.

He has pointed out that the 19-century British heyday of Dickens and Austen, for example, saw the publication of perhaps 20,000 or 30,000 novels—the huge majority of which are never studied.

The problem with this "great unread" is that no human can sift through it all. "It just puts out of work most of the tools that we have developed in, what, 150 years of literary theory and criticism," Mr. Moretti says. "We have to replace them with something else."

Something else, to him, means methods from linguistics and statistical analysis. His Stanford team takes the Hardys and the Austens, the Thackerays and the Trollopes, and tosses their masterpieces into a database that contains hundreds of lesser novels. Then they cast giant digital nets into that megapot of words, trawling around like intelligence agents hunting for patterns in the chatter of terrorists.

Learning the algorithms that stitch together those nets is not typically part of an undergraduate English education, as several grad students point out over pastries in the lab one recent morning.

"It's hard to teach English Ph.D. students how to code," says Kathryn VanArendonk, 25, a ponytailed Victorianist whose remark draws knowing chuckles from others.

But the hardest thing to program is themselves. Most aren't trained to think like scientists. A control group? To study *novels*? How do you come up with pointed research questions? And how do you know if you've got valid evidence to make a claim?

One of the more interesting claims the group is working on is about how novels evolved over the 19th century from preachy tales that told readers how to behave to stories that conveyed ideas by showing action. On a whiteboard, Long Le-Khac, 26, sketches how their computational tools can spit out evidence for the change: the decline of abstract conceptual words like "integrity," "loyalty," "truthfulness." Mr. Jockers chimes in with the "So what?" point behind this chart: The data are important because scholars can use these macro trends to pinpoint evolutionary mutants like Sir Walter Scott.

"It's very tantalizing to think that you could study an author effect," Mr. Jockers says. "So that there's this author who comes on the scene and does something, and that perpetuates these ripples in the pond of prose."

What they have right now is more like a teaspoon of prose. To achieve what they really want—the ability to make generalizations about all of literature without generalizing, because they are supported by data—what they need is a much larger archive.

An archive like Google's.

The Library

If Google Books is like a haphazardly organized used-book shop, as one university provost has described it, Daniel J. Clancy is its suitably rumpled proprietor.

The freckled former leader of an information-technology research organization at NASA is now engineering director of Google Books. He works a few miles down the road from Mr. Jockers on a surreal corporate campus that feels like it was designed by students high on LSD: lava lamps, pool tables, massage parlors, balloons, gourmet grub, a British-style red phone booth, doors that lead nowhere, and rafters hung with a toy snake. A proposed settlement he negotiated with authors and publishers would permit the use of millions of in-copyright works owned by universities for "nonconsumptive" computational research, meaning large-scale data analysis that is not focused on reading texts. Mr. Clancy would turn over the keys to his bookshop, plus $5-million, to one or two centers created for this work—the centers that Stanford and others hope to host.

"It's pretty simple," he says. "We'll give them all the books."

Mr. Clancy says this with the no-big-deal breeziness of someone who works for a Silicon Valley empire of nearly 21,000 employees, one whose products creep into every media business. But for scholars, those three words—"all the books"—are a new world.

The digital content available to them until now has been hit or miss, and usually miss, says John M. Unsworth, dean of the Graduate School of Library and Information Science at the University of Illinois, one of the partners in the HathiTrust consortium. Google has changed the landscape. Pouring hundreds of millions into digitization, the company did in a few years what Mr. Unsworth believes would have taken libraries decades: It has digitized over 12 million books in over 300 languages, more than 10 percent of all the books printed since Gutenberg.

"We haven't had digitized collections at a scale that would really encourage people broadly—across literary studies and English, say—to pick up computational methods and grapple with collections in new ways," Mr. Unsworth

says. "And we're about to now."

But here's the rub. Google Books, as others point out, wasn't really built for research. It was built to create more content to sell ads against. And it was built thinking that people would read one book at a time.

That means Google Books didn't come with the interfaces scholars need for vast data manipulation. And it isn't marked with rigorous metadata, a term for information about each book, like author, date, and genre.

Back in August 2009, Geoffrey Nunberg, a linguist who teaches at the University of California at Berkeley's School of Information, wrote an article for *The Chronicle* that declared Google's metadata a "train wreck." The tags remain a "mess" today, he says. When scholars start trying large-scale projects on Google Books, he predicts, they'll have to engage in lots of hand-checking and hand-correction of the results, "because you can't trust these things."

Classification is particularly awful, he adds. A book's type—fiction, reference, etc.—is key information for a scholar like Mr. Jockers, who can't track changes in fiction if he doesn't know which books are novels. "The average book before 1970 at Google Books is misclassified," Mr. Nunberg says.

Mr. Clancy counters that Google has made "a ton of progress" improving the data, a claim backed up by Jean-Baptiste Michel, a Harvard systems-biology graduate student with intensive experience using the corpus for research. Mr. Clancy also points out that the metadata come from libraries and reflect the quality of those sources. Many of the problems always existed, he says. It's just that people didn't know they existed, because they didn't have Google's full text search to find the mislabeled books in the first place.

And Google is finally opening its virtual stacks to digital humanists, with a new research program whose grant winners are expected to be announced by the end of May.

But don't expect text-mining to sweep the humanities overnight. Or possibly ever.

"There are still a tremendous number of historians, for example, that are really doing very traditional history and will be," says Clifford A. Lynch, director of the Coalition for Networked information. "What you may very well see is that this becomes a more commonly accepted tool but not necessarily the center of the work of many people."

A Clash of Methodologies

As the humanities struggle with financial stress and waning student interest, some worry that the lure of money and technology will increasingly push computation front and center.

Katie Trumpener, a professor of comparative literature and English at Yale University who has jousted with Mr. Moretti in the journal *Critical Inquiry,* considers the Stanford scholar a deservedly influential original thinker. But what happens when his "dullard" descendants take up "distant reading" for their research?

"If the whole field did that, that would be a disaster," she says, one that could yield a slew of insignificant numbers with "jumped-up claims about what they mean."

Novels are deeply specific, she argues, and the field has traditionally valued brilliant interpreters who create complex arguments about how that specificity works. When you treat novels as statistics, she says, the results can be misleading,

because the reality of what you might include as a novel or what constitutes a genre is more slippery than a crude numerical picture can portray.

And then there's the question of whether transferring the lab model to a discipline like literary studies really works. Ms. Trumpener is dubious. Twenty postdocs carrying out one person's vision? She fears an "academia on autopilot," generating lots of research "without necessarily sharp critical intelligences guiding every phase of it."

Her skepticism is nothing new for the mavericks in Mr. Moretti's lab. When presenting work, they often face the same question: "What does this tell me that what we can't already do?"

Their answer is that computers won't destroy interpretation. They'll ground it in a new type of evidence.

Still, sitting in his darkened office, Mr. Moretti is humble enough to admit those "cellars of culture" could contain nothing but duller, blander, stupider examples of what we already know. He throws up his hands. "It's an interesting moment of truth for me," he says.

Mr. Jockers is less modest. In the lab, as the day winds down and chatter turns to what might be the next hot trend in literary studies, he taps his laptop and jackhammers his knee up and down. "We're it," he says.