Digital Humanities/Lecture 10/ May 5th 2023

What Is Distant Reading?

By Kathryn Schulz
The New York Times, June 24, 2011

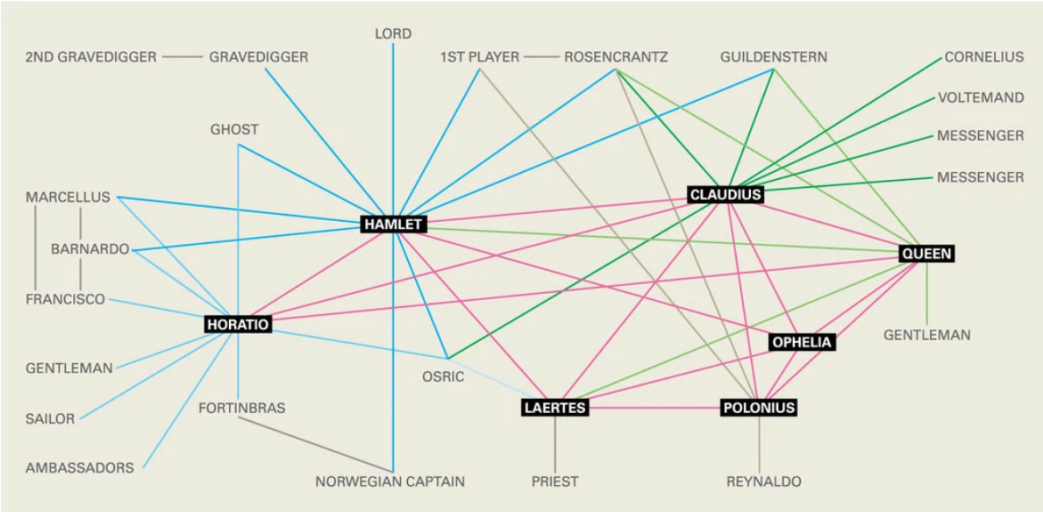

Illustration by Joon Mo Kang (Source: Stanford Literary Lab)

"Ars longa," the ancient saying goes, "vita brevis." Art is long, life short, and the problem is intensifying. As the literary ars lurches exponentially more longa — accommodating the printing press, "Gravity's Rainbow," Google Books — our collective TBR [to be read] pile towers ever more vertiginously overhead. Which raises a question: What are we mortal beings supposed to do with all these books?

Franco Moretti has a solution: don't read them. Moretti is not a satirist. He's an Italian literary scholar and the founder of the Stanford Literary Lab, which opened last year, published its maiden pamphlet in January and followed up with another last month. The first pamphlet asks whether computers can recognize literary genres, and the second uses network theory to re-envision plots.

As its name suggests, the Lit Lab tackles literary problems by scientific means: hypothesis-testing, computational modeling, quantitative analysis. Similar efforts are currently proliferating under the broad rubric of "digital humanities," but Moretti's approach is among the more radical. He advocates what he terms "distant reading": understanding literature not by studying particular texts, but by aggregating and analyzing massive amounts of data.

We need distant reading, Moretti argues, because its opposite, close reading, can't uncover the true scope and nature of literature. Let's say you pick up a copy of "Jude the Obscure," become obsessed with Victorian fiction and somehow manage

to make your way through all 200-odd books generally considered part of that canon. Moretti would say: So what? As many as 60,000 other novels were published in 19th-century England — to mention nothing of other times and places. You might know your George Eliot from your George Meredith, but you won't have learned anything meaningful about literature, because your sample size is absurdly small. Since no feasible amount of reading can fix that, what's called for is a change not in scale but in strategy. To understand literature, Moretti argues, we must stop reading books.

The Lit Lab seeks to put this controversial theory into practice (or, more aptly, this practice into practice, since distant reading is less a theory than a method). In its January pamphlet, for instance, the team fed 30 novels identified by genre into two computer programs, which were then asked to recognize the genre of six additional works. Both programs succeeded — one using grammatical and semantic signals, the other using word frequency. At first glance, that's only medium-interesting, since people can do this, too; computers pass the genre test, but fail the "So what?" test. It turns out, though, that people and computers identify genres via very different features. People recognize, say, Gothic literature based on castles, revenants, brooding atmospheres, and the greater frequency of words like "tremble" and "ruin." Computers recognize Gothic literature based on the greater frequency of words like . . . "the." Now, that's interesting. It suggests that genres "possess distinctive features at every possible scale of analysis." More important for the Lit Lab, it suggests that there are formal aspects of literature that people, unaided, cannot detect.

The lab's newest paper seeks to detect these hidden aspects in plots (primarily in Hamlet) by transforming them into networks. To do so, Moretti, the sole author, turns characters into nodes ("vertices" in network theory) and their verbal exchanges into connections ("edges"). A lot goes by the wayside in this transformation, including the content of those exchanges and all of Hamlet's soliloquies (i.e., all interior experience); the plot, so to speak, thins. But Moretti claims his networks "make visible specific 'regions' within the plot" and enable experimentation. (What happens to Hamlet if you remove Horatio?)

Some insights do emerge from this paper's 57 diagrams, as when the nascent divide between court and state in Renaissance Europe becomes visible in the network. Reading the paper, though, I mostly vacillated between two reactions: "Huh?" and "Duh!" — sometimes in response to a single sentence. For example, Moretti, quoting a colleague, defines "protagonist" as "the character that minimized the sum of the distances to all other vertices." Huh? O.K., he means the protagonist is the character with the smallest average degree of separation from the others, "the center of the network." So guess who's the protagonist of Hamlet? Right: Hamlet. Duh.

Distant reading might prove to be a powerful tool for studying literature, and I'm intrigued by some of the lab's other projects, from analyzing the evolution of chapter breaks to quantifying the difference between Irish and English prose styles. But whatever's happening in this paper is neither powerful nor distant. (The plot networks were assembled by hand; try doing that without reading Hamlet.) By the end, even Moretti concedes that things didn't unfold as planned. Somewhere along the line, he writes, he "drifted from quantification to the qualitative analysis of plot."

I admire Moretti's honesty in saying this: most scholars, whatever their disciplinary background, do not publish negative results. But I would admire it more if he didn't elsewhere dismiss qualitative literary analysis as "a theological exercise." (Moretti does not subscribe to literary-analytic pluralism: he has suggested that distant reading should supplant, not supplement, close reading.) The counterpoint to theology is science, and reading Moretti, it's impossible not to notice him jockeying for scientific status. He appears now as literature's Linnaeus (taxonomizing a vast new trove of data), now as Vesalius (exposing its essential skeleton), now as Galileo (revealing and reordering the universe of books), now as Darwin (seeking "a law of literary evolution").

The trouble is that Moretti isn't studying a science. Literature is an artificial universe, and the written word, unlike the natural world, can't be counted on to obey a set of laws. Indeed, Moretti often mistakes metaphor for fact. Those "skeletons" he perceives inside stories are as imposed as exposed; and literary evolution, unlike the biological kind, is largely an analogy. (As the author and critic Elif Batuman pointed out in an n+1 essay on Moretti's earlier work, books actually are the result of intelligent design.)

But Moretti isn't interested in the unquantifiable, inscrutable actions of intelligent human beings trying to write stuff. There will always be some people for whom new technologies seem to promise completeness and certainty, and Moretti, enthusing over the prospect of "a unified theory of plot and style," is one of them. Literature, he argues, is "a collective system that should be grasped as such." But this, too, is a theology of sorts — if not the claim that literature is a system, at least the conviction that we can find meaning only in its totality.

Moreover, as theologies go, Moretti's is neither new nor, at present, rare. The idea that truth can best be revealed through quantitative models dates back to the development of statistics (and boasts a less-than-benign legacy). And the idea that data is gold waiting to be mined; that all entities (including people) are best understood as nodes in a network; that things are at their clearest when they are least particular, most interchangeable, most aggregated — well, perhaps that is not the theology of the average lit department (yet). But it is surely the theology of the 21st century.