Literary **Lab**

Literature, Measured

Franco Moretti

Pamphlets of the Stanford Literary **Lab**

Franco Moretti

Literature, Measured[1]

Pamphlet One. In 2010, none of the five authors of "Quantitative Formalism" had any idea they were writing a "pamphlet". A well-known scholarly journal had been asking for an article on new critical approaches, and that's where we sent the piece once it was finished. But it came back with so many requests for corrections, that it felt like a straightforward rejection. It was dismaying; a few years ago, computational criticism was still shunned by the academic world, and we couldn't help thinking that what was being turned down was not just an article, but a whole critical perspective. And since we also thought that the essay was perfectly fine as it was, we decided that – instead of trying our fortune with another journal (or, god forbid, make the required alterations) – we would publish it on our own, as a document of the Literary Lab. I cannot remember how the term "pamphlet" came up; and, frankly, it wasn't even right: pamphlets have a public vocation that our work, with its heavily technical aspects, couldn't possibly have. But the word captured the euphoria of being on our own; the freedom to publish what we wanted, when and how we wanted: short, long, even *very* long, our pamphlets never come out a minute earlier than they're ready, nor a minute later, either; and without going through the grinder of editing "styles". And all this, because "Quantitative Formalism" was rejected by —. Never mind. They did us a favor.

Meandering. "Nothing is as rare as a plan", Napoleon is supposed to have said, and we have certainly proved him right. We never know where the next pamphlet will come from: they have ranged from the individual research of undergraduate theses ("Loudness in the Novel") and Ph.D. dissertations ("Becoming Yourself: the Afterlife of Reception"), to the elective affinity of a pair of like-minded researchers ("A Quantitative Literary History of 2,958

---

1 The pages that follow were written as the introduction to a French collection of Literary Lab pamphlets, edited by Jérôme David (ITAC, Paris, 2016).

Nineteenth-Century Novels: the Semantic Cohort Method"; "Between Canon and Corpus: Six Perspectives on 20th-Century Novels"; "Bankspeak: the Language of World Bank Reports, 1946-2012"), and the complex polyphony of larger groups ("Quantitative Formalism"; "Style at the Scale of the Sentence"; "On Paragraphs. Scale, Themes, and Narrative Form"; "Canon/ Archive. Large-Scale Dynamics in the Literary Field"; plus several ongoing projects). With time, however, a team of five or six researchers has emerged as the most frequent formation, and the one that best embodies the novelty of laboratory work. Take "Style at the Scale of the Sentence". It began with the six of us dividing the initial tasks among ourselves; we promptly disagreed on the path to follow; research options were opened, then abandoned; there were two or three collective presentations at the Lab in the first year; in between, bursts of solitary work, small-group discussions, and rivers of emails; later, a long coda of drafts, discussions, and re-formulations. The last few months were crucial for the final section of the pamphlet, which converted two years of empirical findings into theoretical conjectures. "Without the concepts of the second half", we wrote, "the results of the first would have remained blind; and without the empirical content of the first part, the categories of the second would have remained empty. Only from their encounter did critical knowledge arise" (P5/28).[2] At the time we didn't know it, but we had just put into words that back-and-forth between the empirical and the conceptual which would characterize all our future research.

Work. Laboratorium, laborare, labour. That's what a lab is: a place to do work. Group work, mostly; and now that I have experienced it in a variety of forms (small and large groups, students, faculty, from the same discipline and university or from different ones), I would say that almost every project goes through two very different stages. In the initial phase, the group functions like a single organism, where every individual attends to a specific function. The first of them, programming: something Matt Jockers laid the foundations for even before the Lab was officially opened, and Ryan Heuser sustained over the years with his unique imaginative talent, and whose mathematical implications have eventually been made clear to us all by Mark Algee-Hewitt. On the basis of programming, much more becomes possible: from the refinement of the corpus to the analysis of initial results; from the review of the critical literature to the design of follow-up experiments. This functional division of labor, whose results no individual scholar could ever achieve in isolation, is clearly indispensable to modern research. But the second stage of group work is, if possible, even better. Now, the team sits together around a table – the lab table: as essential a tool as the really expensive ones – and discusses how to make sense of the results. Here, the efficient integration of the first stage gives way to a swirl of disparate associations: C reflects on the language of a specific excerpt, and A on the historical categories that could explain it; F recalls something D had said a few months earlier (and then forgotten); E recognizes a grammatical pattern, for which B suggests an evolutionary explanation... All researchers bring to this phase their interests, and even fixations. At times, there is a lot of noise. But in a few magic moments,

---

2 From now on, pamphlets will be quoted directly in the text, indicating the pamphlet number first, and the page number after the diagonal slash. All pamphlets can be downloaded from the Literary Lab website.

the group becomes truly more than the sum of its parts; it "sees" things that no single pair of eyes could have. If, in the pamphlets that follow, there are some genuine discoveries, that's where they have always come from.

Adagietto. A scientific essay, composed like a Mahler symphony: discordant registers that barely manage to coexist; a forward movement endlessly diverted; the easiest of melodies, followed by leaps into the unknown. I have often tried to write like this, and always failed. Then, with the pamphlets, the form has suddenly emerged. It unfolds along four distinct, nearly equivalent levels: images; captions; text; and footnotes. Images, first of all: time plots, histograms, trees, networks, diagrams, scatterplots... Images come first, in our pamphlets, because – by visualizing empirical findings – they constitute *the specific object of study of computational criticism;* they are our "text"; the counterpart to what a well-defined excerpt is to close reading. Next to them, and equally new, captions: almost absent from the early pamphlets, captions have since become as essential for our work as descriptions are in art history, or observations in scientific reports; writing them has taught us to observe more attentively, and to declare what we "see" in any given image, thus announcing what the first steps of the analysis will be. After images and captions, the main body of the text: challenged and compressed by the two newcomers, the text is forced to become tighter, sharper: it must weave the four registers into a single argument without robbing them of their new-found autonomy; it must be narrative *and* theoretical; take a sequence of discrete events, and transform them into a conceptual grid; and all of this, in fewer pages than most academic articles! It's hard; but it's good for the clarity of the argument. And then, if one really needs more space, there are always the footnotes: the space where we put our bibliographical cards on the table, examine theoretical alternatives, and fantasize on future studies; a mix of homework, polemics, and speculation which adds its own subterranean twist to the complexity of the whole. To do justice to this heterogeneous architecture – to make it *visible* – our designer Jake Coolidge has invented a "page" (for "Canon/Archive") made of three large columns, where the four registers take turns in unpredictable combinations. It's strange, and wonderful, and a little unsettling, seeing your thinking mirrored so closely in the shape of the page. At times, more then Mahler, it feels like *Tristram Shandy*.

Past and present. In the early 1990s, as I was trying to put together a collective *Atlas of World Literature,* Fred Jameson suggested that keeping a log-book of the whole enterprise would be a valuable document in its own right. As funding agencies turned down the *Atlas,* that original log-book never materialized; but Jameson's idea resurfaced, twenty years later, in the detailed reporting which is typical of our pamphlets. Reporting is the right form for the "exploratory" aspect of digital humanities work: it keeps track of our forays into the immensity of the new digital archive, anchoring them to solid facts: such and such data *have* been found, and here they are, black on white. And then, recording the doubts and the decisions of the research process makes you fully understand what you have done. It's not by accident that many of our pages are initially drafted in the past tense, and then revised into the present: it's never quite clear whether the main point lies in relating a specific moment in the

unfolding of past research, or presenting a thesis for present discussion. And it's not clear, because most of the time it's both. A section of "Style at the Scale of the Sentence" begins by explaining how we came to correlate verb forms and novelistic genres, and in the space of four or five paragraphs has turned into a discussion of how the *Bildungsroman* shaped the idea of modern youth (P5/18-25). Empirical findings, and conceptual work, again. But, as we will see next, the process is not always so virtuous.

"Till a man is sure he is infallible..." In the shift from report to reflection, a very special role is played by – failures. In the early days of experimental science, write Shapin and Shaffer, Robert Boyle considered it "necessary [...] to offer readers circumstantial accounts of failed experiments", because they showed that he "was not willfully suppressing inconvenient evidence", thus assuring his readers "that he was such a man as should be believed."[3] Now, there is no doubt that the reporting of failures has something to do with persuasion; missteps make for an entertaining narrative, and self-criticism is always a good way to pre-empt the criticism of others. Yet, the main reason to report failures has nothing to do with capturing the benevolence of readers; it's that failures *throw a unique light on the whole research process.* Failures take us all the way back to our starting points: to those unspoken assumptions that go "without saying", and thus easily escape critical scrutiny. Looking for the semantic center of tragic form in moments of maximum conflict, for instance – a "failure" discussed at length in "Operationalizing" (P6/9-13) – rested on the assumption that, in drama, words and deeds are in synchrony with each other; discovering that this was *not* the case showed that the theory of tragedy needed a fundamental re-thinking. Looking for style at the scale of the paragraph – and, again, not finding it (P10/4-5) – convinced us that a single top-to-bottom theory of the literary text (like that of stylistics) couldn't possibly be right, and opened the way to new hypotheses on textual scale. "What may produce a decisive advance", wrote Popper in *The Logic of Scientific Discovery,* is more often than not "the modification of what we are inclined to regard [because of its complete agreement with our normal habits of thought] as obviously innocuous".[4] Exactly. By frustrating our expectations, failed experiments "estrange" our natural habits of thought, offering us a chance to transform them. Or as Boyle had memorably put it four centuries ago: "till a man is sure he is infallible, it is not fit for him to be unalterable".[5]

So what? There comes a moment, in digital humanities talks, when someone raises the hand and says: "Ok. Interesting. But is it really *new?*" Good question... And let's leave aside the obvious lines of defense, such as "but the field is still only at its beginning!", or "and traditional literary criticism, is *that* always new?" All true, and all irrelevant; because the digital

---

3 Steven Shapin and Simon Shaffer, *Leviathan and the Air-Pump. Hobbes, Boyle, and the Experimental Life*, Princeton UP, 1985, pp. 64-5.

4 Karl Popper, *The Logic of Scientific Discovery*, 1934, Harper and Row, New York 1965, p. 76.

5 Robert Boyle, "A Proemial Essay, wherein, with some Considerations touching Experimental Essays in general, Is interwoven such an Introduction to all those written by the Author, as is necessary to be perused for the better understanding of them"; in *The Works of the Honourable Robert Boyle*, ed. Thomas Birch, 2. ed., London: J.&F. Rivington, 1772. vol. I, p. 311.

humanities have presented themselves as a radical break with the past, and must therefore produce evidence of such a break. And the evidence, let's be frank, is not strong. What is there, moreover, comes in a variety of forms, beginning with the slightly paradoxical fact that, in a new approach, not everything has to be new. When "Network Theory, Plot Analysis" pointed out, in passing, that a network of *Hamlet* had Hamlet at its center (P2/4), the *New York Times* gleefully mentioned the passage as an unmistakable sign of stupidity. Maybe; but the point, of course, was not to present Hamlet's centrality as a surprise; it was exactly the opposite: had the new approach *not* found Hamlet at the center of the play, its plausibility would have disintegrated. Before *using* network theory for dramatic analysis, I had to *test* it, and prove that it corroborated the main results of previous research. Corroboration, alas, is often boring to humanities scholars (and clever journalists); but it *has* long played a role in scientific research, and having introduced it into our field is an achievement, not a weakness of the digital humanities. Besides, seldom is corroboration *just* corroboration. By and large, "On Paragraphs" validates a thematic approach to literature; but in doing so we also discovered that themes have an elective affinity with the scale of the paragraph (P10/5-8); that the typical paragraph has, not one, but between two and four distinct themes (P10/13-15); that the connection between thematics and narratology rests precisely on this plural-yet-limited number (P10/16-17); and so on, from specification to specification. Thematics has not been revolutionized, but it has certainly been changed, and, why not, improved. It's the "encounter between concepts and measurements" (P5/28) of "Style at the Scale of the Sentence"; a few months later, in "Operationalizing", the encounter had become a "radicalization of our relationship to concepts" (P6/13); radicalization, in the sense that – when you have to turn a concept into a series of operations – you look at it in an analytical fashion, that opens the way to its critique. I am thinking of how the concept of "scale" changed across three of our pamphlets, for instance: still largely a metaphor in the "mortar, bricks, and architecture" of "Quantitative Formalism" (P1/8), "scale" found a solid textual anchor in the sentence of "Style at the Scale of the Sentence", and was then generalized in the refrain of "different scales, different features" that concluded "On Paragraphs" (P10/21-22).[6] It's fascinating, how a series of quantitative measurements enters into a dialogue with concepts, and slowly transforms them. *Slowly*. Forget the hype about computation making everything faster. Yes, data are gathered and analyzed with amazing speed; but the *explanation* of those results – unless you're happy with the first commonplace that crosses your mind – is a different story; here, only patience will do. For rapidity, nothing beats traditional interpretation: Verne's "Nautilus" means – childhood; Count Dracula – monopoly capital. One second, and everything changes. In the lab, it takes months of work.

Beyond concepts? After corroboration and conceptual revision, a few results that seem to be genuinely new. First of all, the "loudness" of Holst Katsma's pamphlet: a notion that,

---

6 Comparable developments have occurred for the notion of "dramatic networks" (from "Network Theory" to "Operationalizing", and the forthcoming "Modeling Dramatic Networks"), and for the conceptual pair which first appeared in "Becoming Yourself" as "consumption" and "conversation" (P3/2ff and 18ff), and returned as "popularity" and "prestige" in "Between Canon and Corpus" (P8/9ff), and then, more analytically, in "Canon/Archive" (P11/3-5).

despite all the interest in "voices" and "polyphony", had never become part of the theory of the novel. Katsma found a way of operationalizing his intuition, and of articulating it into a semantics and a grammar of loudness; he supported it with readings of *Pride and Prejudice* and *The Idiot*, and concluded with the discovery of the "quieting down of the English novel" (P7/19-25), which he followed in its decade-by-decade unfolding. "Bankspeak" also focused on a gradual historical process – the transformation of the language of the World Bank's yearly reports – and in fact took precisely the *slowness* of this evolution as typical of how an "institutional style", with its complex of rules and constraints, eventually crystallizes (P9/13-17). (And now, it's impossible not to day-dream of a follow-up study in which slowness becomes the key tempo of all that is "institutional" in literary life – genres, styles, movements, canons… – thus also transforming our idea of literary history as a whole.) And finally, after loudness and slowness, a novelty which is not a concept, but – a script: "Correlator". Developed by Ryan Heuser and Long Le-Khac in order to find "the words that most closely correlate with a given 'seed' word" in terms of historical frequency, Correlator generated large "word cohorts", allowing Heuser and Le-Khac to discover unexpected long-term trends (P4/6-7). We have here a typical product of what could be called "the programming imagination"; a form of thinking *that fuses together the formulation and the operationalization of concepts,* leaving them often half-implicit *as concepts*, while liberating their full force *as algorithms*. People of an older generation may miss the clarity of categorical distinction; but in the years to come, the main contribution of computational criticism to literary study may well come from these Centaur-like creatures, half script, and half theory. Correlator, is the harbinger of a new species.

Triangulations. As concepts moved increasingly to the foreground of our work, so did our engagement with existing theories. Three distinct intellectual areas have been particularly significant in this respect. The first is the great formalist tradition, from Russian Formalists to Spitzer's and Auerbach's stylistics, some aspects of structuralism, and recent work in corpus linguistics. This lineage is the one we are closest to in terms of objects and categories ("morphology", "genre", "register", "system", "style"); and, since form is the repeatable element of literature, this is also where we turn to in order to set the process of quantification in motion; in "Canon/Archive", for instance, we operationalized the difference between canon and archive, not via their respective semantic content, but in terms of formal features like redundancy and lexical variety (P11/5ff). After this first lineage, so clearly hyper-literary, comes one that is not literary at all: the epistemology of the natural sciences, taken in a rather broad sense. Here, we have found our inspiration a little haphazardly, borrowing principal component analysis from genetics, for instance, and network theory from mathematics and physics, and entropy from information theory; not to mention specific concepts like "measurement" (Kuhn), "instrument" (Koyré), or "normal/pathological" (Canguilhem), all of which have played a role in this or that pamphlet. That it all required a double work of translation – from natural to literary objects, and from concepts to algorithms – became itself part of the point: it defined our view of the digital humanities as *the form taken by a scientific-ex-*

planatory approach in the digital age: triangulating Canguilhem with charts of type-token ratio and forgotten Victorian novels, or Koyré with network statistics and minor characters in plays – *this* is how we understand the new approach. Finally, third major presence, Bourdieu. Evoked, for one reason or another, in "Quantitative Formalism", "Becoming Yourself", "Style", "Between Canon and Corpus", and engaged at length in "Canon/Archive" (P11/3-5, 12), "Bourdieu" stands for a literary study that is empirical and sociological at once. Which, of course, is obvious. But he also stands for something less obvious, and rather perplexing: the near-absence from digital humanities, and from our own work as well, of that other sociological approach that is Marxist criticism (Raymond Williams, in "A Quantitative Literary History", being the lone exception). This disjunction – perfectly mutual, as the indifference of Marxist criticism is only shaken by its occasional salvo against digital humanities as an accessory to the corporate attack on the university – is puzzling, considering the vast social horizon which digital archives could open to historical materialism, and the critical depth which the latter could inject into the "programming imagination". It's a strange state of affairs; and it's not clear what, if anything, may eventually change it. For now, let's just acknowledge that this is how things stand; and that – for the present writer – something needs to be done. It would be nice if, one day, big data could lead us back to big questions.