FRANCO MORETTI & OLEG SOBCHUK

HIDDEN IN PLAIN SIGHT

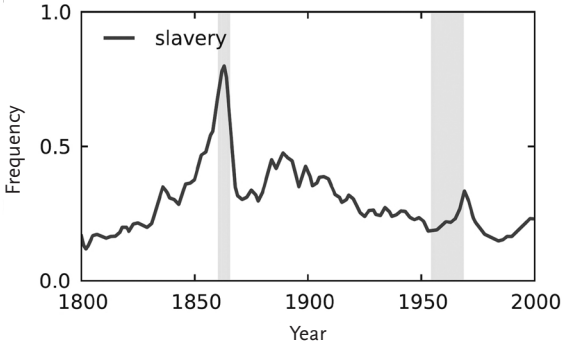
Data Visualization in the Humanities

If there is one feature that immediately distinguishes the digital humanities (DH) from the 'other' humanities, data visualization has to be it. Histograms, scatterplots, time series, diagrams, networks . . . ten, fifteen years ago, studies of film, music, literature or art didn't use any of these. Now they do, and here we examine some premises (unspoken, and often probably unconscious) of this field-defining practice. Field-defining, because visualization is never *just* visualization: it involves the formation of corpora, the definition of data, their elaboration, and often some sort of preliminary interpretation as well. Whence the idea of this article: to gather sixty-odd studies that have had a significant impact on DH, and analyse how they visually present their data.[1] What interests us is visualization *as a practice*, in the conviction that practices—what we learn to do by doing, by professional habit, without being fully aware of what we are doing—often have larger theoretical implications than theoretical statements themselves. Whether this has indeed been the case for DH, is for readers to decide.[2]

I. HISTORY

We begin with the article that announced the creation of Google Ngrams, thus catapulting the digital-quantitative approach into the open, well beyond the boundaries of a small academic niche: 'Quantitative Analysis of Culture Using Millions of Digitized Books', published in *Science* in January 2011. Figure 1, opposite, is the first image one encounters in the article, and it sets the tone for all that follows: the horizontal axis measures the passage of time; the vertical one, the frequency of the word 'slavery'.

Figure 1: *Quantitative Analysis of Culture Using Millions of Digitized Books*, 2011



'It was not until the middle of the eighteenth century that a common visual vocabulary for time maps caught on. But the new linear formats of the eighteenth century were so quickly accepted that, within decades, it was hard to remember a time when they were not already in use. The key problem in chronographies, it turned out, was . . . how to create a visual scheme to clearly communicate the uniformity, directionality, and irreversibility of historical time.'

Daniel Rosenberg and Anthony Grafton, *Cartographies of Time*

---

[1] The corpus for our meta-analysis was based on several criteria. First, we have focused on those humanities disciplines that have only recently turned to quantification, excluding both linguistics and social history, where the quantitative turn occurred much earlier. We have also excluded papers dedicated to a single text or author, and those of which we were single authors (though several articles of which we are co-authors are criticized in the pages that follow). With these criteria in mind, we went through all major literary and DH journals, starting from 2010. Of the resulting 62 papers, about half are directly addressed in this article (and listed in its final endnote), while the other half can be found on the *New Left Review* website.

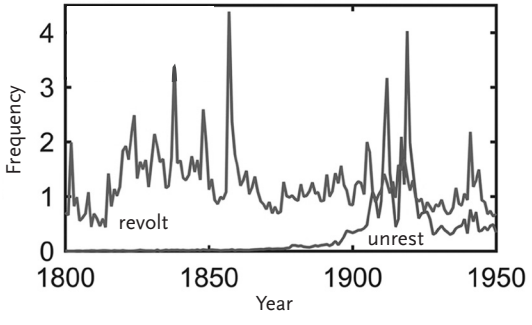
We have tried to combine several fields of study—but we both work mostly on literature, and the greatest part of our evidence will come from there, thus limiting the validity of some of our claims. Furthermore, our criteria are clearly questionable, and different ones would have produced a different sample and overall assessment. Finally, after this article had been completed, in January 2019, another survey of DH scholarship—Nan Z. Da's 'The Computational Case Against Computational Literary Studies'—was published in the spring 2019 issue of *Critical Inquiry*. Since the aims and methods of the two articles are fundamentally different, we have decided to leave our text unchanged.

[2] For the record, we do think this is the case: as the humanities were taken completely by surprise by the sudden availability of digital archives and computational systems, the primacy of practice over theory was probably almost inevitable. Cultural history was not expecting these novelties, and, more importantly, was not *in need* of them, in the way early modern astronomy—to resort to a parallel that has often been evoked—had developed a theoretical need for something like the telescope. Given this starting point, it's not surprising that the concrete use of the new tools—the practice—preceded and overshadowed their theoretical justification.

A time series, as this type of chart is usually called: the years pass, and the frequency of 'slavery' changes; it doubles around the Civil War, it slowly declines to its initial frequency, it rises again, more modestly, at the time of the civil rights movement, and so on. 'Quantitative Analysis of Culture' includes 33 charts, and 27 of them—80 per cent—are of this kind.

Though 80 per cent is high for our corpus, time series are unquestionably very common in DH work, and have thus become its visual 'signature'.[3] Simplicity, as Rosenberg and Grafton suggest, has certainly helped. Just two elements: history and semantics. One word (Figure 1), two (Figure 2), four (Figure 3), or hundreds of them, as in the 'semantic fields' and 'topics' of Figures 4 and 5. The numbers change, and so do the objects under investigation (books, newspaper articles, World Bank reports, novels, scholarly studies); what doesn't change is the focus on content. 'Topic modelling'; 'content analysis'; 'text mining': meaning is like a raw material, unaffected by textual organization. Corpora are 'bags of words', as the saying has it; meaning must be extracted—text *mining*—and that's it: once out, it's perfectly explicit: 'Changes in discourse *reveal* broader historical and sociocultural changes . . .'; 'The models . . . *reveal* a strong decline of positive emotionality through time . . .'; 'This approach *reveals* important but hitherto unarticulated trends'. Language reveals; it never hides, or lies, or complicates matters. It's an idea of culture, as the triumph of the explicit.

FIGURE 2: *Content Analysis of 150 Years of British Periodicals*, 2017



[3] In literary study, for instance, they are never absent from those articles that—by appearing in 'core' disciplinary journals like *Modern Language Quarterly*, *Poetics Today*, or *New Literary History*—have acted as a bridge between the old and the new approach.

Figure 3: *Bankspeak: The Language of World Bank Reports*, 2015

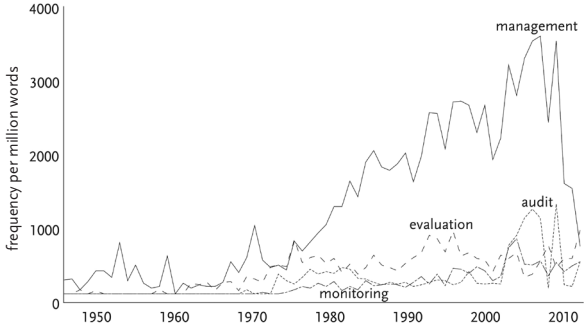


Figure 4: *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*, 2012

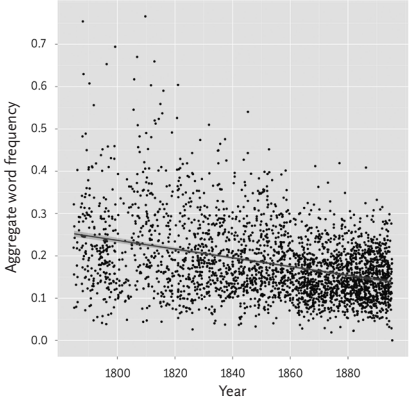
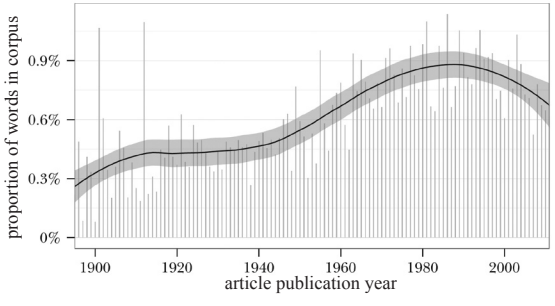


Figure 5: *The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us*, 2014



More on this later. Now, shifting from the vertical to the horizontal axis, it's striking how often these time series extend over a historical span of exactly a century. The novels and scholarly articles of Figures 4 and 5, the lexicon of property in parliamentary debates (Figure 6), bestsellers written by women (Figure 7), shot length in film (Figure 8), the expression of emotions in fiction (Figure 9), contractions in American novels (Figure 10), repetitions in the canon and the archive (Figure 11), reviews of poetry collections (Figure 12) . . . Topic after topic, the century has emerged as the typical yardstick of quantitative cultural history.

FIGURE 6: *Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora*, 2018

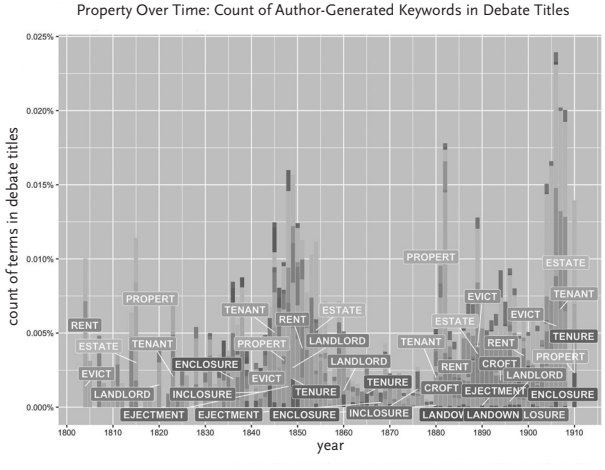


FIGURE 7: *The Transformation of Gender in English-Language Fiction*, 2018

FIGURE 8: *Shot Durations, Shot Classes and the Increased Pace of Popular Movies*, 2015

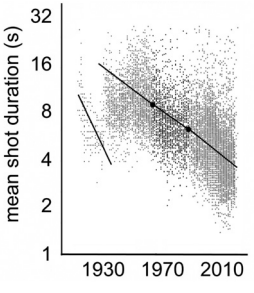


Figure 9: *Birth of the Cool: A Two-Centuries Decline in Emotional Expression in Anglophone Fiction*, 2016

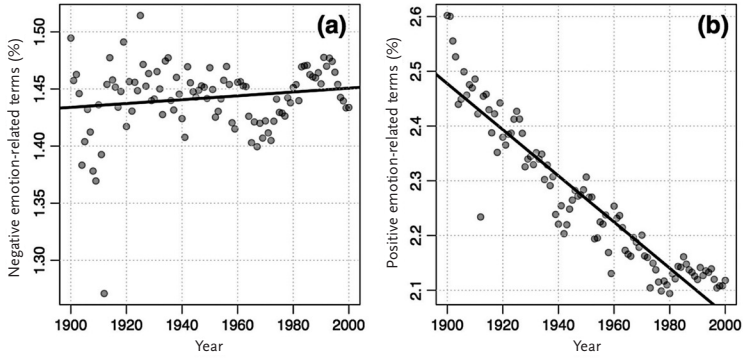


Figure 10: *The Making of Middle American Style*, 2016

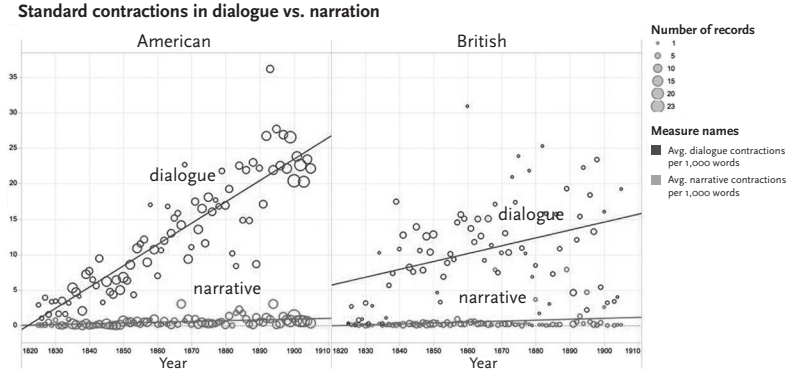


Figure 11: *Canon/Archive: Large-Scale Dynamics in the Literary Field*, 2016

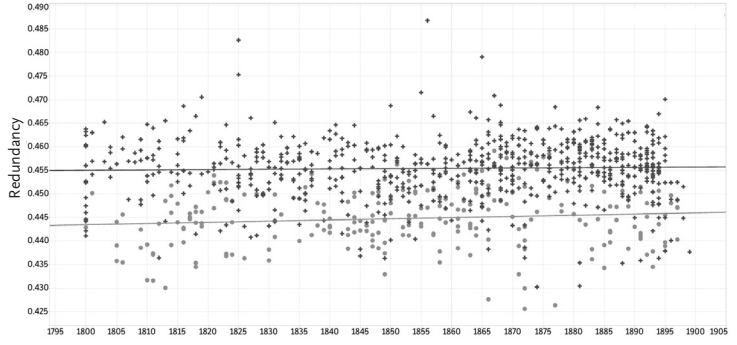


Figure 12: *The* Longue Durée *of Literary Prestige*, 2016

In a few cases, it's a matter of external constraints: film has existed for about a hundred years, and there is nothing one can do about that; books published between 1800 and 1900 allow good optical recognition (unlike earlier ones), while being free from copyright restrictions (unlike later ones)—whence our typical over-production of nineteenth-century studies. But the deeper reason for this predominance of the century has probably to do with DH's claim to be 'a way of discovering and interpreting patterns *on a different historical scale*' ('The Quiet Transformations of Literary Studies'). Different from previous research, often limited to a narrow historical span. But different *how*, exactly?

In searching for an answer, the century offered itself as an option so obvious, it went almost without saying. Intuitively, a century is long; it's not anthropomorphic (as we seldom live that long), and is therefore extraneous to the old focus on individual authors; it is the tempo *of the world*, not of life. Romance languages use it for informal periodization (*El Siglo de Oro, une dix-neuvièmiste, il Quattrocento*); American universities, for many of their hires. The notion was there, in the existing *doxa*; a nice round number, long enough to suggest a new dimension, but not so long as to be unmanageable. True, it wasn't really *a concept*; it had no place, for instance, in the tripartition of historical time—*longue durée*, cycle, event—elaborated by the *Annales* school; and it certainly wasn't adopted as a result of a theoretical decision. But practice trumped theory, once again: the century offered an intuitive frame for the new

scale we were after, and we turned it into the pedestal—the horizontal axis—for our historical findings.

A keyword

'A different historical scale': specifically, one *in which trends become visible*. Up to a few years ago, no one spoke of trends in the humanities; you couldn't, as long as you studied only a few texts, spanning a handful of years. With centuries, you can. New fields need keywords, and 'trend'— with its mix of direction and measurement—was perfect for DH; not by accident, it showed up right away, in the abstract of that 2011 article in *Science* ('Analysis of this corpus enables us to investigate cultural trends . . .'), and returned as a sort of opening chord in the first page of studies on film shots, jazz evolution, literary scholarship, British periodicals, poetry reviews, legal records, and expressions of emotion; in a single article ('Quantitative Literary History of 2,958 Novels'), it occurs sixty-eight times. And it's not just a word: in Figures 4, 8, 9, 10, 11, 12 and 13 trends are *physically present* in the form of regression lines and analogous elaborations of the data. We haven't just talked of trends; we have made them *visible*, and given them pride of place.

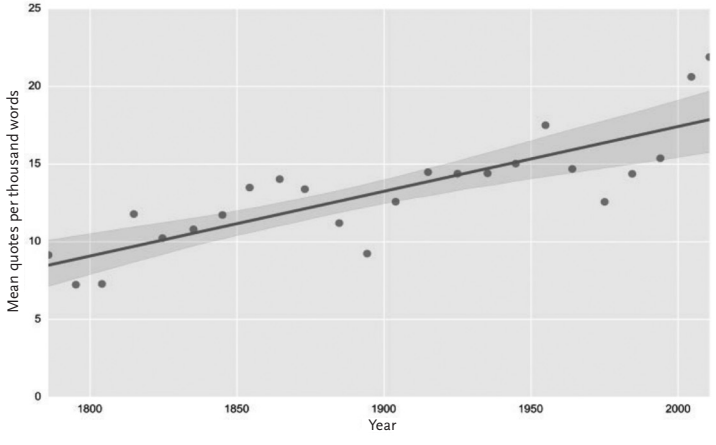
Oxford English Dictionary: 'Trend': 'The general direction which a stream or current, a coast, mountain-range, valley, stratum, etc. tends to take.' This, near the end of the eighteenth century. (Before, the word had already existed for a long time—since the eleventh century, apparently—as a verb: a river trends; a coastline trends). Then, at the end of the nineteenth century, the figurative meaning emerges: a discussion trends: 'it turns in some direction'; it 'has a general tendency'; 'the general course, tendency, or drift (of action, thought, etc.)' Course, tendency, drift, direction . . . always in the singular, because thinking in trends means *reducing the many to one*: a cloud of data, to a single line. Whereas previous attempts tried to visualize the 'many intersecting trajectories of history', write Rosenberg and Grafton in *Cartographies of Time*, 'the form of the timeline . . . emphasized overarching patterns and the big story.'[4] From multiple trajectories to a single big story: this is the key. Data are always confusing and full of noise: trend lines are perfectly univocal. They make data easy to read; they give them *a meaning*.

[4] Daniel Rosenberg and Anthony Grafton, *Cartographies of Time*, New York 2010, p. 20.

Irresistible temptation, finding a meaning in history. Figure 13: the presence of dialogue in a corpus of English-language novels. The chart 'shows that [writers] add roughly one quote per thousand words every 25 years, equivalent to just under one more quote per page every century . . .' Actually, the chart shows that writers double the number of quotes in the first half century, then oscillate inconclusively for about 50 years, remain mostly stable for the following 70–80, and then, rather suddenly, increase them by 50 per cent in the last two decades. (Or perhaps: they generate two 80-year cycles where a considerable rise ends in a steep final decline, with a third cycle having just started in the 1980s.) Writers do all sorts of things, *except* adding one quote every 25 years: it's the trend line that does that.

Or take Figure 8, above, on the average shot length in film between 1913 and 2013. 'The decline is not abrupt nor is it demonstrably articulated at any point', state the authors; 'it is uniform and gradual over the course of at least eighty years'. Not abrupt, when silent films more than halved shot length in fifteen years? Uniform over eighty years (1930–2010), when duration clearly *increased* between 1930 and 1960? Don't they see their own data? Of course they do; but trend lines have changed *how* we see: they have transformed statistical abstractions into physical presences as real as the data themselves—and in fact, usually, *far more*

Figure 13: *Dialogism in the Novel: A Computational Model of the Dialogic Nature of Narration and Quotations*, 2017



visible than the data themselves. Visualization appeals to our intuition; if it shows a cloud of dots with a line in the middle, we only look at the line. It's inevitable. And so, instead of helping us analyse the evidence, averages have often allowed us to forget it. We turned to quantification because we wanted to see all those documents that the predominance of the canon had made invisible—and now that they are in front of our eyes, we have found a way not to see them![5]

Time series; content; centuries; trends. One step at a time, a new kind of cultural history has emerged from the practice of visualization. And so have its polemical targets. The opening paragraph of 'The Quiet Transformations of Literary Studies':

> The history of literary study is primarily remembered as a narrative of *conflicting ideas*. Critical movements *clash* . . . Although scholars have complicated this . . . with an emphasis on *social and institutional struggle*, *generational conflict* remains a central framework: instead of *struggles* among ideas there are *struggles* among genteel amateurs, professionalized scholars, and so on. In emphasizing *conflict*, these approaches still leave aside important dimensions of the history of scholarship: assumptions that change quietly, without explicit debate; entrenched patterns that survive the visible *conflicts*; long-term transformations of the terrain caused by social change [emphasis added] . . .

A few more lines, and the authors mention 'the century-long trend' of their first chart (and of many that follow): trends, replacing the old 'narrative of conflicting ideas' as the central mechanism of history. (Having appeared 8 times in the first 14 lines of the article, 'conflict', 'struggle' and 'clash' return only 6 more times in the following 25 pages; 'century-long trends', for their part, appear in 21 of the 23 charts.) And indeed, it's hard to think of struggles in front of a trend line—and quite easy, by contrast, to envision a 'quiet transformation' of the historical landscape. And so, though probably no one wanted this to happen, our exaggerated

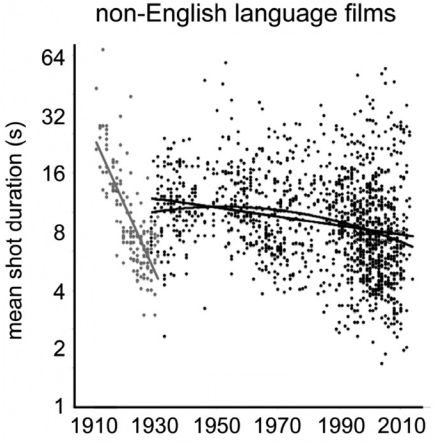
[5] Though canons and averages are clearly not the same thing, they play a comparable role in the drastic simplification of the cultural field: 'averages', wrote Ernst Mayr in 1959, 'are merely statistical abstractions: only the individuals of which the populations are composed have reality'. See 'Darwin and the Evolutionary Theory in Biology', in B. J. Meggers, ed., *Evolution and Anthropology: A Centennial Appraisal*, Washington, DC 1959, p. 29.

reliance on trends has *de facto* banished conflict from DH research, creating an odd, vaguely disheartening 'There Is No Alternative' atmosphere. Cultural history deserves better than that.

Scala della ragione

There is another problem, with trends. While claiming that shot length declined uniformly as a result of 'more intensified demands on viewers' attention', the authors of 'Shot Durations, Shot Classes' also report that a sampling of three genres showed that 'the mean shot duration for . . . action films [was] 4.64 seconds, and that for . . . comedies and dramas 8.55 seconds'. The figures make sense: a genre dominated by rapid physical movements encourages shorter shots than genres that rely on sustained dialogue. So one returns to Figure 8, and wonders: could the decline in shot length be mostly due to *the increased presence of action films in the American film industry*? Since we are not told about the genre composition of the study's corpus, it's a plausible hypothesis, made even more so by the chart of *non*-American films in Figure 14: where—action films being certainly less numerous than in the US—shot length changes very little between 1930 and 2010. If length depended on 'intensified attention', the results should be the same everywhere (short of claiming that US audiences are the most attentive of all); since the results are in

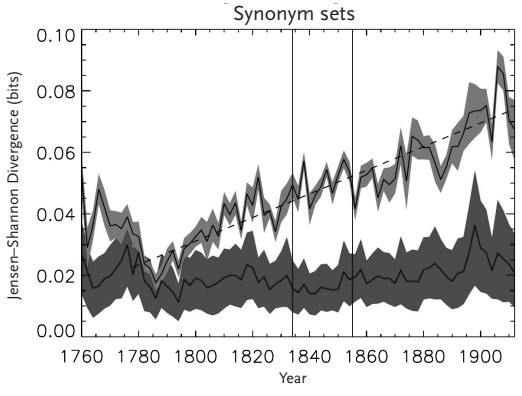
FIGURE 14: *Shot Durations, Shot Classes and the Increased Pace of Popular Movies*, 2015



fact very different, the appeal of action films—with their brutal narrative simplification—may well be the reason for the overall transformation.[6] Instead of a trend with a single there-is-no-alternative direction—and of the optimistic diagnosis of an ever-increasing attention—we find ourselves looking at a polarization of the film industry in opposite directions: at a conflict of forces, ultimately. It's a different way of thinking about historical change.

One last point. At times, trends are unquestionably there—and they're indeed quite uniform and gradual. But this doesn't necessarily mean that *the forces behind them* are equally uniform. Take 'The Civilizing Process in London's Old Bailey': a study of 160 years of legal records that charts the growing differentiation of the language used for violent and non-violent crimes (Figure 15). There are local oscillations, here, but the trend is as regular as one could wish. And yet, as the authors point out, this 'secular rise . . . is not simply the amplification of a particular initial pattern; rather, the actual synonym sets that serve to distinguish violent from non-violent trials are themselves changing.' What this means is made clear by

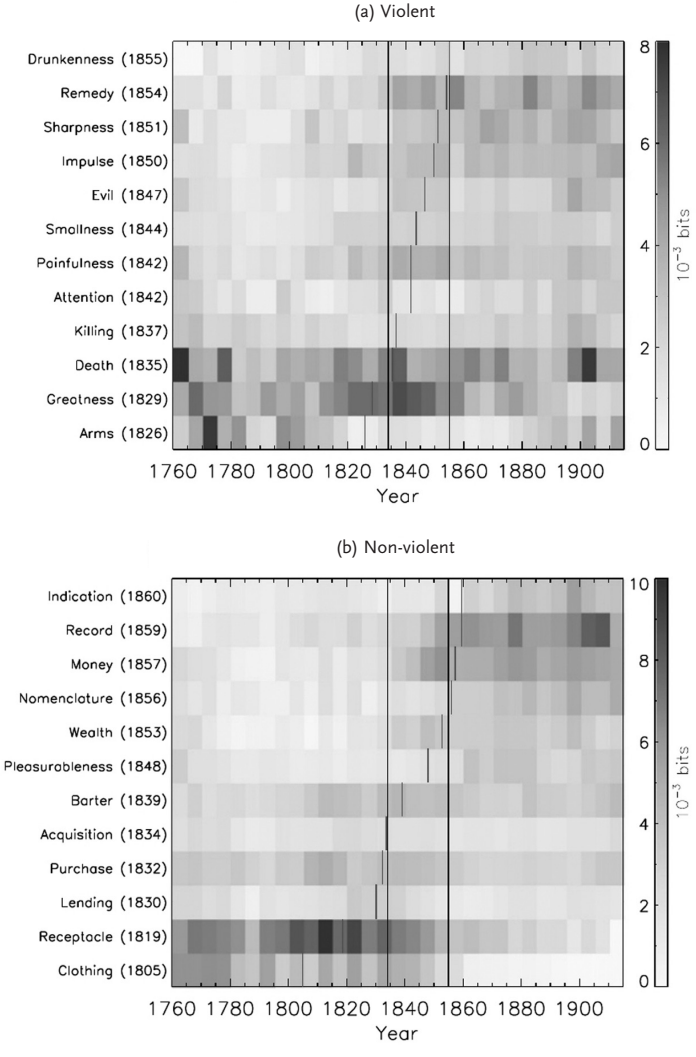
FIGURE 15: *The Civilizing Process in London's Old Bailey*, 2014



[6] The same is true of a parallel study—'Quicker, Faster, Darker: Changes in Hollywood Film Over 75 Years' (2011)—which finds 'a gradual, essentially linear change over 75 years' in the 'visual activity index' of film. Given that the index for *Toy Story* is 15 times higher than that for *Barry Lyndon*, and that films like *Toy Story* have become immensely more frequent than those like *Barry Lyndon* in the past several decades, it's quite likely that the increase in children's films is the best explanation for the change under discussion.

Figure 16, which lists the decade-by-decade frequency of the lexicon used for the study. Here, nothing is regular. Some sets, like the synonyms of 'death', remain almost unchanged throughout the period; 'arms' and 'greatness' peak in the first half of the nineteenth century, and 'remedy' (a sign of medical evidence entering the picture) in the second half; in the non-violent group, the 'receptacle' and 'clothing' of petty thefts are important early on, and then disappear; the opposite holds for 'money'

Figure 16: *The Civilizing Process in London's Old Bailey*, 2014



and 'record', harbingers of Victorian white-collar crime. The trend was gradual and uniform; the forces that produced it, disparate and heterogeneous. 'The Making of Middle American Style': a constant, steady rise of the colloquial style over a century and more (see above, Figure 10); behind it, five separate waves, each lasting roughly the space of a generation: from Irish, Scottish and English archaisms ('o', 'thee', 'thy'), to polite forms of address ('sir', 'doctor'), African-American and working-class terms ('fer', 'ter', 'tuh', 'dat'), and the first-name basis of twentieth-century dialogue (on this, see the complementary study, 'Operationalizing the Colloquial Style: Repetition in 19th-Century American Fiction', 2017). The rise was perfectly regular; its generation-by-generation phases, entirely contingent. One changes historical scale, and finds a different picture.

An analogy will help explain what we are trying to say. In several Italian cities, one encounters in the old part of town a Palazzo della Ragione—the palace of reason that, in early modern times, used to house law courts, notaries, magistrates and branches of government. In Verona, in the mid-fifteenth century, a Scala della Ragione was added to the Palazzo, in the Cortile del Mercato Vecchio: a splendid red marble staircase, that turns sharply at a 90° angle in mid-ascent, yet preserves a steady and constant incline: exactly the way reason should work. But this beautiful regularity rests on four wildly dissimilar arches (Figure 17)—just as the gradual trend of the Old Bailey records depended on a mosaic of discordant and uneven forces. And one wonders: what should historical explanation be like, in such cases? Should we focus on the uniform long-term trend—or on the strange underlying arches? And if

FIGURE 17: *Verona, Scala della Ragione,* c. 1450

the latter, does this mean that the trend is a mere surface effect, a sort of statistical phantom? Or should our conceptual architecture try to resemble the Scala, rectilinear and dissonant at once? But what would that even mean, for historical categorization . . .

II. MORPHOLOGY

We have seen DH time series clustering around the century, or thereabouts. In the visualization of morphological features, however—the stream of consciousness, the rhythm of American situation comedies, the role of the protagonist in drama, the direction of the gaze in European portraits—the historical range becomes far more varied: 30 years, 60, 350, 500 . . .[7] And in studies that dispense with time series altogether, variation fluctuates even more, from the 3 years of popular songs and the 10 of British 'theological novels', to the 2,000 and more of Aby Warburg's *Pathosformeln*, and the wholly indeterminate spectrum—'at the intersection of oral folkloric narratives and literary styles and contexts'—of European fairy tales.[8]

Why this insouciance with history? Because the object of study has changed. In describing the difference between the 'evolutionary' and the 'functional' biologist, Ernst Mayr observed how the latter's focus on 'the operation and interaction of structural elements' created a situation in which those 'elements' are not followed over time, as would happen in an evolutionary study, but are rather immobilized in order to 'eliminate, or control, all variables'.[9] 'The word "morphology" means the study of the component parts'—wrote Vladimir Propp in the opening page of the *Morphology of the Folktale*—'in their relationship to each other and to the whole.'[10] For this type of study, more than history one needs a dissecting table.

---

[7] 'Turbulent Flow: A Computational Model of World Literature' (2016); 'Distant Viewing: Analysing Large Visual Corpora' (2018: under review); 'Distributed Character: Quantitative Models of the English Stage, 1550–1900' (2017); 'How Portraits Turned Their Eyes Upon Us: Visual Preferences and Demographic Change in Cultural Evolution' (2013).

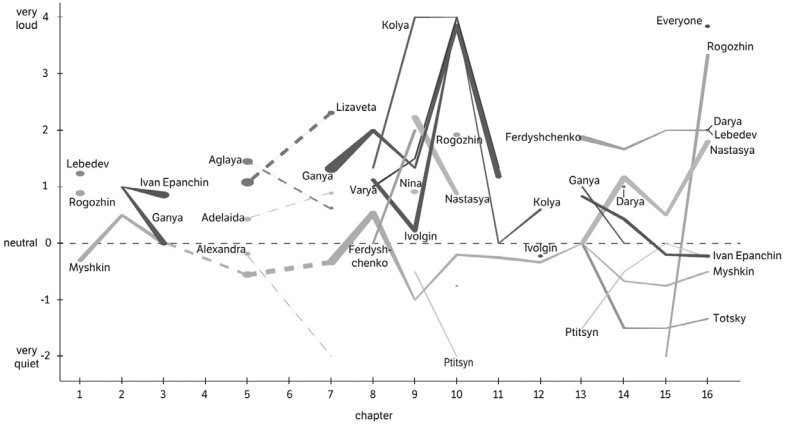
[8] 'Are Atypical Things More Popular?' (2018); 'Advances in the Visualization of Data: The Network of Genre in the Victorian Periodical Press' (2015); 'Totentanz: Operationalizing Aby Warburg's *Pathosformeln*' (2017); 'Computational Analysis of the Body in European Fairy Tales' (2013).

[9] Ernst Mayr, *Toward a New Philosophy of Biology*, Cambridge, MA 1988, p. 25.

[10] Vladimir Propp, *Morphology of the Folktale*, Austin 1968 [1927], p. xxv.

We have moved full circle from the beginning of this article. In time series, texts were viewed as 'bags of words'—with zero consideration for 'structural elements' or 'relationships to the whole'—that directly 'revealed' the surrounding world. By contrast, morphology concentrates on the 'operations and interactions' of structures, placing them *between* the observer and the world. Look at the chart of loudness in Dostoevsky's *The Idiot* in Figure 18: whereas the first dozen images of this article didn't include a single hint on how texts worked, here we see *nothing but* the internal organization of the novel. The author of the study—a pianist, as well as a literature major—was interested in the 'volume' of narrative texts, and found an elegant way of operationalizing it by parsing speaking verbs into loud, neutral and quiet.[11] He also analysed 'what' characters said, of course, but, mostly, *how* they said it: from his chart, you can tell that at the end of the sequence Rogozhin shouts and Myshkin whispers—but you have no idea what they are saying. But this bold reduction made it possible to transform Bakhtin's metaphor of

FIGURE 18: *Loudness in the Novel*, 2014



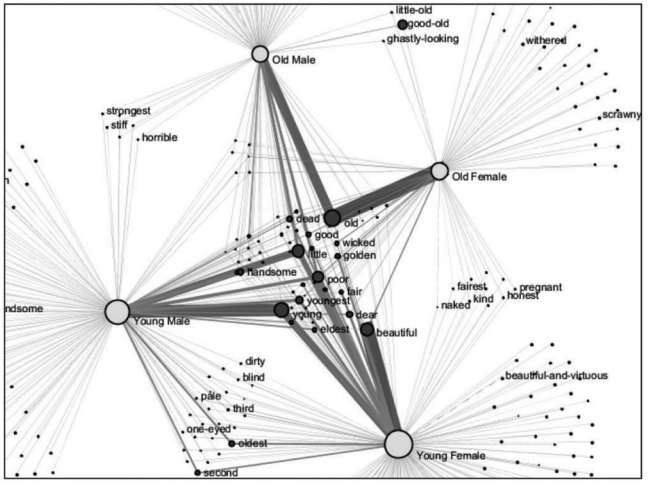
'Though the loudness is created by many discrete individuals, [this image] is not a graph of cacophony. Emerging from the many vocal lines, one can see . . . a slight peak in chapter 7; second, the dramatic peak in chapter 10; third, a bifurcation of dialogue into the extremes: Rogozhin, Lebedev, Ferdyshchenko, Darya, Nastasya . . . filling the room with loudness; Myshkin, Ivan Epanchin, Ptitsyn, and Totsky creating an undercurrent of whispers. [This] is truly a graph of polyphony, of voices "artistically organized".'

[11] Three examples from a famous scene of *Alice in Wonderland* explain the principle of classification followed in the essay. Loud register: '"Off with their heads!" shouted the Queen.' Neutral: '"I suppose so", said Alice.' Quiet: 'He whispered, "She's under sentence of execution."'

novelistic polyphony into a concrete, reliable analytical strategy. *The Idiot* had been turned into a score.

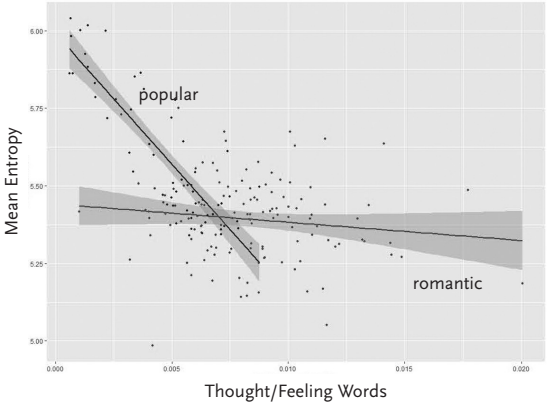
'The relationship of component parts to each other', read Propp's definition of morphological analysis. In 'Loudness in the Novel', the component parts are voice and plot; in another study, 'Computational Analysis of the Body in European Fairy Tales' (2013), they are a matrix of four key characters (young/old, male/female), and of all the adjectives that define them in the corpus (Figure 19). The thickness of the lines is proportional to the frequency of the association, and it shows that—in the given corpus—male and female are less in opposition than young and old: young male and female are solidly linked to each other, as are the two old characters—whereas young and old remain quite disconnected. In comparable fashion, Figure 20 presents the correlation between interiority and linguistic repetition in East Asian literature; and Figure 21 (overleaf) that between syntax and semantics in nineteenth-century English novels.

Figure 19: *The Body in European Fairy Tales*, 2013



Adjectives in the middle of the star-like structure, like 'poor' and 'good', are common to three, or to all four characters; others are shared by only two of them (like the 'small' that joins the two young characters, or the 'ghastly-looking' of the two old ones); while others still ('scrawny' for the old female, 'beautiful-and-virtuous' for the young female) apply to one character only. Taken together, the adjectives exemplify the texture of value judgments that underlie this corpus of fairy tales.

Figure 20: *Self-Repetition and East Asian Literary Modernity 1900–30*, 2018



'Plots for the ratio of "thought/feeling" words against average entropy for Japan and China, with linear regression lines fitted by genre. In both cases, we can observe that as the ratio of "thought/feeling" words increases (horizontal axis), the mean entropy of the texts decreases (vertical axis), indicating more lexical repetition.'

The association between interiority and repetition emerging from this chart of Chinese novels is far from obvious: in the same years, the inner landscape of Ulysses's *stream of consciousness rested on the very opposite principle—a complete* absence *of repetition and predictability.*

Time without history

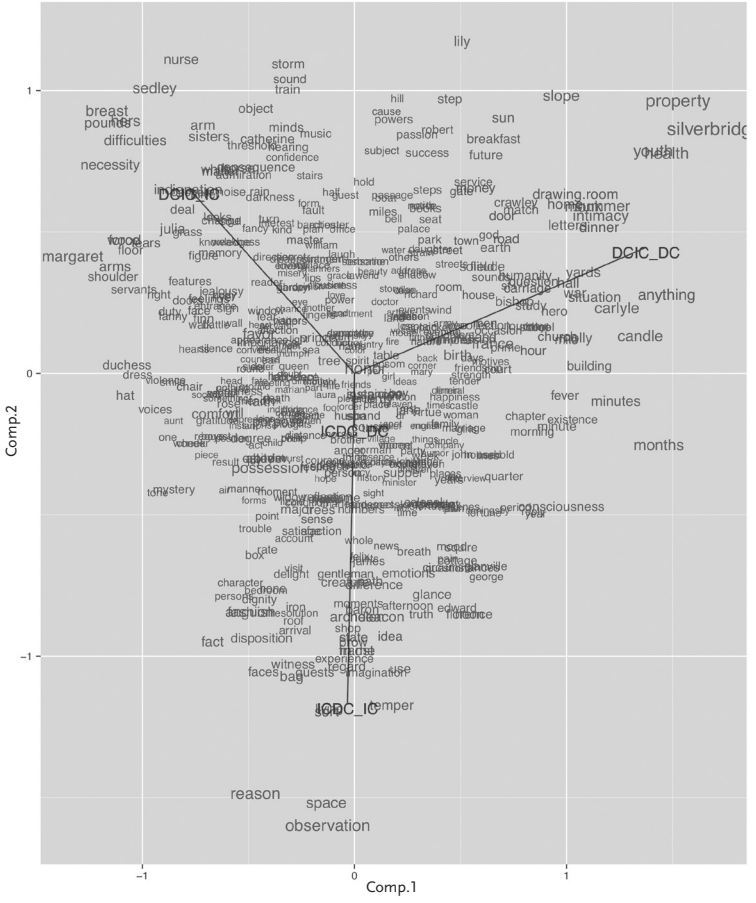
'Unlike what can be observed in other fields of history', Ernest Labrousse once remarked, 'in economic history all that matters is repeated.'[12] And in morphology, too: its favourite objects—patterns—emerge precisely from the regular reiteration of the same process over time.[13] But this 'time' is

[12] Labrousse is quoted in Krzysztof Pomian's retrospective reflection on the *Annales*, where he also points out that, with Braudel, 'the study of repetitions transcends the realm where it appeared to be confined . . . and ends up occupying almost the entire horizon of the historian': 'L'histoire des structures', in Jacques Le Goff, ed., *La nouvelle histoire*, Paris 2006 [1978], pp. 119, 121. To have enlarged the horizon of the cultural historian in a similar fashion has been a major achievement of DH; when it comes to the novelty of the results and the clarity of concepts, though, we still have a lot to learn from what the *Annales* group managed to do—with far smaller archives, and more primitive computational tools—one, two generations ago.

[13] Patterns have played a similar role within morphology to that of trends in history: they are often greeted as a sufficient result *in themselves*, whereas the real challenge lies in discovering their underlying causes. For a critical discussion, see 'Patterns and Interpretation', in Franco Moretti, ed., *Canon/Archive: Studies in Quantitative Formalism From the Stanford Literary Lab*, New York 2017.

very different from that of the economic historian, or of the time series we encountered earlier. The 250 novels in Figure 21 had all been published in the nineteenth century, and it would have been easy to import their publication dates into the image; but we were too intent on understanding the internal mechanisms of those sentences to even *think* about

FIGURE 21: *Style at the Scale of the Sentence*, 2013



An example of what literature has become in the new space of literary labs, where novels are 'prepared' for analysis in a way that severs all connections with the lived experience of reading them. In this image, sentences and words are extracted from their narrative context and recombined in wholly abstract fashion to test the possible relationship between syntactic sequences and semantic fields.

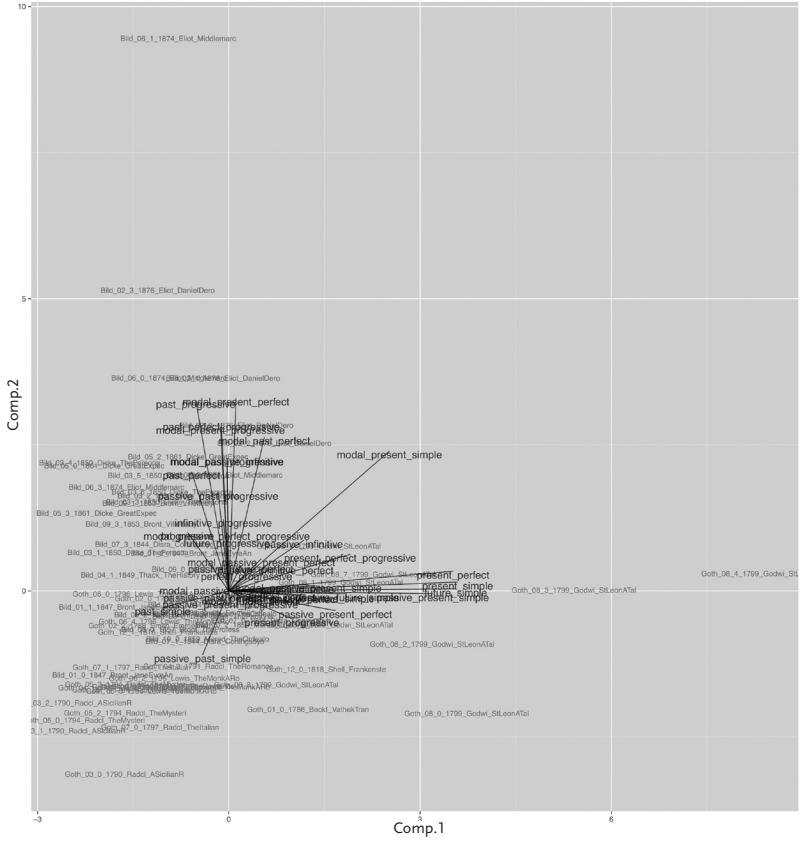
history, let alone do something about it. Our study needed 'time', in the sense that only repetition could establish the correlations we were looking for; but it wasn't the time of the historian: it was *the time of the lab*. Abstract time: a few variables, re-run hundreds of times to understand 'operations and interactions'. The time of the experiment, to be kept rigorously *separate* from the time of the world.[14] Time, without history.

In the next section we will see that morphology is not inevitably imprisoned in the abstract time of the experiment; for now, let's just observe how the computational turn has simultaneously offered a great opportunity to morphological study, and magnified its problems. An opportunity, because the operationalization of aesthetic categories has made them more tangible than ever before: 'polyphony', '*Pathosformeln*', 'minor characters'—nowadays, these abstractions can literally be *seen*, and (some of) their components accurately measured. But computation has also magnified the problems of morphology, because it has demanded such an artificial isolation of forms, that their historical significance has almost evaporated. Form as the most profoundly social aspect of the work of art: this idea, that seemed so promising half a century ago, has so far eluded computational work, which has been torn between the naive historicism described in the first part of this article, and the abstract, lab-like formalism of the second. An old curse has returned under a new guise.

'The relationship of component parts to each other and to the whole', wrote Propp in the *Morphology*. Figures 18, 19, 20 and 21, with their two variables displayed along the axes of a Cartesian diagram, or distributed over a principal component scatterplot, are good illustrations of the 'relationship of component parts *to each other*'; none of them, however—and in fact, no study that we know of—has ever managed to visualize 'the relationship of component parts *to the whole*'. Some studies have 'condensed' entire forms onto a single feature, and then compared them to each other on this basis—as in the reduction of Gothic novels and *Bildungsroman* to their verb forms in Figure 22 (overleaf), for example; eventually, the basis for comparison has become

[14] 'The chief technique of the functional biologist is the experiment', observed Mayr in *Toward a New Philosophy of Biology*, 'and his approach is essentially the same as that of the physicist and the chemist.'

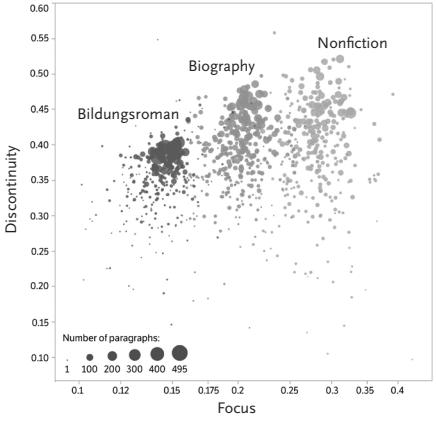
Figure 22: *Style at the Scale of the Sentence*, 2013



'Gothic novels cluster in the lower left quadrant of the chart, and Bildungsroman*, in the quadrant above . . . The units in the chart (Goth_03_0_1790_Radcl_ASicilianR in the bottom left corner, or Bild_06_1_1874_Eliot_Middlemarc in the upper left one) are sections of novels, each containing 200 narrative sentences. The verb forms most responsible for the separation between the two genres are the perfect and the passive past simple for the Gothic, modals and progressives for the* Bildungsroman*.'*

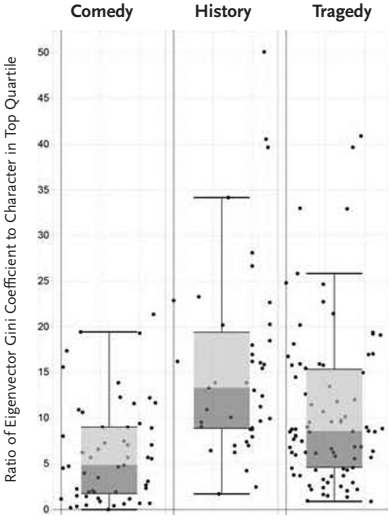
more robust and complex, as in the combination of different metrics into the 'thematic focus' of Figure 23, or of different aspects of network centrality into the 'protagonism' of Figure 24. At bottom, though, these charts have continued to elude the question of how 'component parts' interact to form a complex structure: by reducing the whole to just a few of its parts, they have offered a useful *simplification* of structures, not their analytical blueprints. Visualizing the *Morphology* remains a task for the future.

Figure 23: *On Paragraphs: Scale, Themes and Narrative Form*, 2015



'In this chart, the x-axis measures the focus of paragraphs (that is to say, how much of a given paragraph is devoted to a single topic), and the y-axis their discontinuity (that is to say, the difference between the topics of successive paragraphs). The separation between the three discourses—and especially between fiction and non-narrative nonfiction—is unmistakable.'

Figure 24: *Distributed Character: Quantitative Models of the English Stage, 1550–1900*, 2017



'While history and tragedy exhibit similar morphological relationships between their cores and peripheries, as indicated by their similar protagonism metric, comedies are structurally different, having a much smaller periphery relative to their core.'

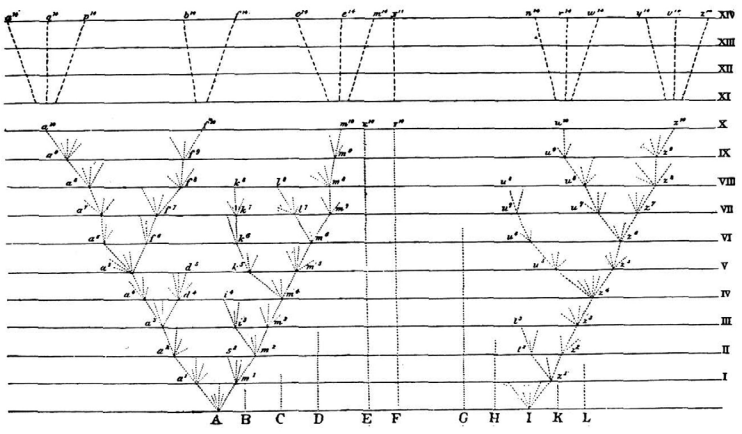
III. HISTORICAL MORPHOLOGY?

A section on history, and one on morphology. Separate, because their aims are different, and in some ways even opposite. Morphology looks for *distinctions and correlations*: features whose interaction gives rise to a characteristic form, all the way to that larger system of distinctions that is a taxonomy. Historical studies aim at *continuity and succession*: instead of opening up their objects, they follow their vicissitudes over time—usually, in the form of a single trend line. And one wonders: what would the combination of these two research strategies look like?

Figure 25 reproduces the only image included in *The Origin of Species*. Inserted in the chapter on 'Natural Selection', the 'diagram' (Darwin's word) aims at making intuitively visible the relationship between the passage of time (measured in thousands of generations along the vertical axis) and the increasing divergence of the initial species A and I indicated at the bottom of the chart. What the diagram shows, in other words, is the inextricable connection of history and morphology that characterizes the natural world. Structures are what they are, because they have *become* that way. There are no forms without history.

No forms without history. But among the hundreds of charts in our corpus, only a handful bear much resemblance to Darwin's diagram.

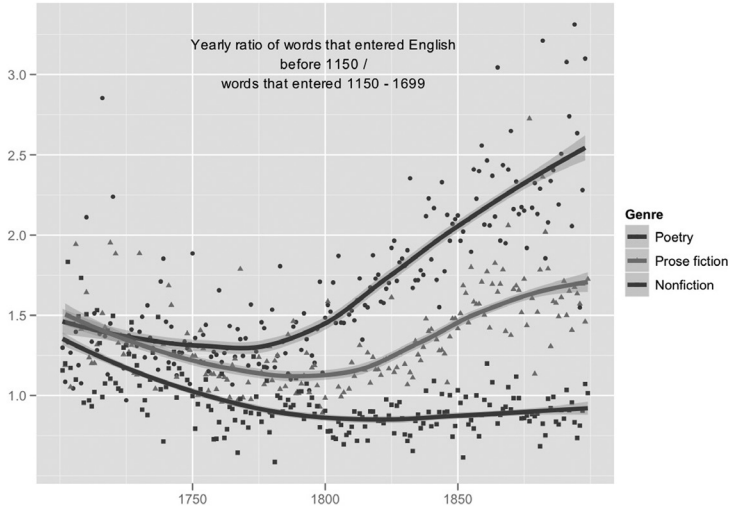
FIGURE 25: *The Origin of Species,* Ch. IV: 'Natural Selection'



'You will find Ch. IV perplexing & unintelligible, without the aid of enclosed queer Diagram, of which I send old & useless proof.'

Charles Darwin, letter to Charles Lyell, 2 September 1859

Figure 26: *The Emergence of Literary Diction*, 2012



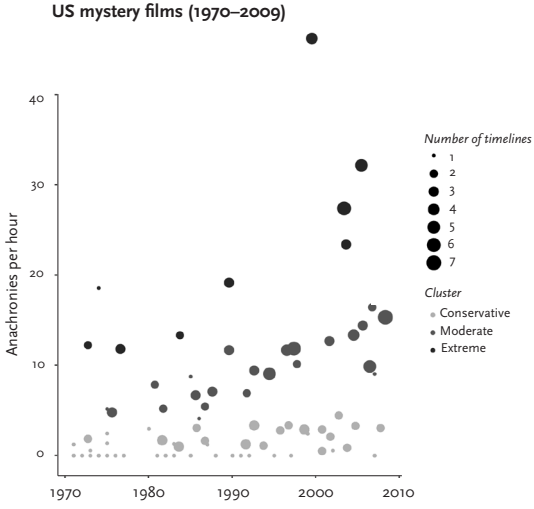
The study measures the yearly ratios between pre-1150 vocabulary and the vocabulary that entered English from 1150 to 1699. This ratio was calculated for more than 4,000 volumes published in the eighteenth and nineteenth centuries.

One of them, reproduced in Figure 26, follows the lexical divergence of English poetry, fiction and nonfiction in the eighteenth and nineteenth centuries. Based on the proportion of words that entered the English language before or after 1150, the chart shows how, between 1750 and 1900, fiction and (especially) poetry increased their use of the pre-1150 vocabulary, whereas nonfiction did not. The resulting pattern recalls Darwin's diagram, as well as the separation of violent and non-violent crimes in the Old Bailey (Figure 15, above), or the branching process that emerges from Figure 27 (overleaf), on flashbacks and flashforwards in film.[15]

[15] A few more articles contain some evidence of branching, but their authors don't pursue the possibility. In 'Quicker, Faster, Darker', for instance, the 'visual activity index' is interpreted as an all-encompassing trend, although the data in one of the article's figures (1b) might indicate an incipient case of divergence, quite similar to that of 'Broken Time'. Something similar happens in 'Film Through the Human Visual System' and in 'Now, Not Now', where bestsellers diverge from prize-winning novels in their use of narrative personas.

Branching processes in the detective fiction of the 1890s, and in free indirect style between 1800 and 2000, had been explicitly discussed in 'The Slaughterhouse of Literature' (2000; now in *Distant Reading*, London and New York 2013) and *Graphs Maps Trees: Abstract Models for Literary History*, London and New York 2005. Neither study possessed however the explicit quantitative dimension we are evaluating here.

FIGURE 27: *Broken Time, Continued Evolution: Anachronies in Contemporary Films*, 2017



In this study, branching was found by chance, as the research team was expecting to see a simple regular increase—a trend: the default assumption of DH *studies—in the use of anachronies (flashbacks and flashforwards). Then it became clear that while* some *films (the dark-grey and mid-grey dots) did in fact show such an increase, another group (the light-grey dots) showed hardly any increase at all. More significantly still, the difference among the three groups wasn't limited to the* number *of anachronies, but included their position within the plot, and, ultimately, their function in the films' narrative structure. In the light-grey group, flashbacks and flashforwards were overwhelmingly located near the beginning or the end of the film, usually with the very explicit function of explaining an enigma via a flashback to the original crime. By contrast, in the dark-grey and mid-grey groups anachronies were distributed rather evenly across films, and seemed to have taken on the different function of multiplying small mysteries everywhere in the story.*

'The Emergence of Literary Diction', 'The Civilizing Process in London's Old Bailey' and 'Broken Time' have three traits in common. In emphasizing the morphological specificity of their objects, they clearly differ from the mainstream of DH historical research; in correlating structural features to the passing of years, they avoid the 'abstract' time of most morphological research; and, finally, they are all inspired by what Ernst Mayr has called 'population thinking'. 'To understand the origin of biodiversity', he wrote, 'it [is] not sufficient to study a single population at different times, so to speak "vertically"; rather, one must compare different contemporary populations of a species with each

other.'[16] Comparing different contemporary populations: instead of looking at a single 'population' of, say, murders, 'The Civilizing Process in London's Old Bailey' follows the simultaneous course of two distinct classes of cases; 'The Emergence of Literary Diction', of three discourse genres; 'Broken Time', of three film forms. By mapping these different 'cultural populations' at regular intervals, a branching pattern takes shape in front of our eyes. It's almost like witnessing the emergence of a new cultural species.[17]

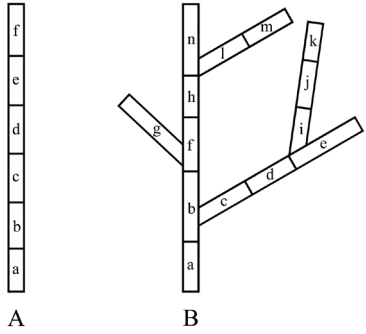
Figure 28 (overleaf), in which Mayr contrasts two different views of evolution, helps understand what is at stake on this point. In phyletic evolution (A), a species evolves over time, moving through various stages—from *a* to *f*—as it adapts to the changing environment; despite its transformations, however, it remains *a single species* throughout. There is no growth of biodiversity. For this 'developmental thinking', as Robert O'Hara has called it, 'history [is] a story of individual development or unfolding—a story of "evolution" in the original sense of the word.'[18] A single thread runs through subsequent stages: *Australopithecus*, *Homo*

[16] Ernst Mayr, *What Evolution Is*, London 2001. 'For those who have accepted population thinking', adds Mayr elsewhere, 'the variation from individual to individual within the population is the reality of nature, whereas the mean value [the "type"] is only a statistical abstraction.' Mayr, *Toward a New Philosophy of Biology*, p. 15.

[17] Needless to say, we are not suggesting that, by combining morphology and history, branching will automatically emerge: Figure 11 above, for instance, which follows canon and archive over a hundred years, shows absolutely no divergence between the two populations. Furthermore, the very opposite of branching—'reticulate' evolution, or the merging of branches—is always a possibility too. The question is addressed in detail in Oleg Sobchuk's *Charting Artistic Evolution: An Essay in Theory*, Tartu 2018. To summarize the central points, an exchange of features across different branches of the tree of culture occurs as a norm among close neighbours, and becomes increasingly unlikely as morphological distance increases. The branch of Gothic novels can quite easily merge with that of historical novels; less easily with courtship novels (though *Northanger Abbey* proves that it's not impossible); even less easily with silver-fork fiction, industrial novels and so on. To paraphrase microbiologist Eugene Koonin: when we look at the evidence from up close we see a reticulate web, but when we look at it from afar we see a tree: 'organisms that appear "close" in phylogenetic trees actually exchange genes frequently, and organisms that seem "distant" in trees are those between which [exchange of genes] is rare.' Eugene Koonin, *The Logic of Chance: The Nature and Origin of Biological Evolution*, Upper Saddle River, NJ 2011, p. 164.

[18] Robert O'Hara, 'Population Thinking and Tree Thinking in Systematics', *Zoologica Scripta*, vol. 26, no. 4, 1997, pp. 325–7.

Figure 28: *'Phyletic Evolution vs. Speciation'*



habilis, *Homo erectus* . . . Classicism, Romanticism, Realism . . . No branches, just a trunk divided into sections—as in the trends we have discussed earlier, where a single line suffices to chart the history of an entire population. In the scenario of speciation on the other hand (B)—in which the ancestor *a* gives rise to the descendant species *g*, *n*, *m*, *k*, *e* . . . —the tree-like shape makes the growth of diversity immediately visible. The populations of poetry and fiction branch off from that of nonfiction; violent trials, off nonviolent ones. History becomes a process of creative diversification.[19]

Tree-like, linear, reticulate . . . why should we even care about the shape of cultural history? We should, because that shape is implicitly a hypothesis about the forces that operate within history; the tentative, intuitive beginning of a theoretical framework. 'Theories are, even more than laboratory instruments, the essential tools of the scientist's trade', wrote Thomas Kuhn over a half century ago;[20] too bad we didn't heed his advice. Although the crass anti-intellectualism of *Wired*—'correlation is enough', 'the scientific method is obsolete'[21]—has fortunately remained an exception, what seems to have happened is that, as the amount of

[19] The argument of this last paragraph returns, from a different angle, to the issue raised earlier in reference to historical trends: in both cases, we object to a kind of visualization that unconsciously assumes a single path as the basic form of historical development.

[20] Thomas Kuhn, 'The Function of Measurement in Modern Physical Science' [1961], in *The Essential Tension: Selected Studies in Scientific Tradition and Change*, Chicago 1977, p. 208.

[21] Chris Anderson, 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired*, 23 June 2008.

quantitative evidence at our disposal was increasing, our attempts at in-depth explanations were losing their strength. Disclaimers, postponements, *ad hoc* reactions, false modesty, leaving inferences 'for another day' . . . such have been, far too often, our inconclusive conclusions.

We are so used to thinking of data and explanations as the two sides of the same coin, that such an outcome seems hard to believe. But it's all around us. The day abundant data will open the way to bold concepts, rather than inhibiting them—that day, the new quantitative cultural history will come into its own, and a real confrontation with the 'other' humanities truly begin.

PAPERS MENTIONED BY YEAR OF PUBLICATION

2011

James E. Cutting et al., 'Quicker, Faster, Darker: Changes in Hollywood Film over 75 Years', *i-Perception,* vol. 2, no. 6

Jean-Baptiste Michel et al., 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science*, vol. 331, no. 6014

2012

Ryan Heuser and Long Le-Khac, 'A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method', *Stanford Literary Lab*, Pamphlet 4

Ted Underwood and Jordan Sellers, 'The Emergence of Literary Diction', *Journal of Digital Humanities*, vol. 1, no. 2

2013

Sarah Allison et al., 'Style at the Scale of the Sentence', *Stanford Literary Lab*, Pamphlet 5

Olivier Morin, 'How Portraits Turned Their Eyes Upon Us: Visual Preferences and Demographic Change in Cultural Evolution', *Evolution and Human Behavior*, vol. 34, no. 3

Scott Weingart and Jeana Jorgensen, 'Computational Analysis of the Body in European Fairy Tales', *Literary and Linguistic Computing*, vol. 28, no. 3

2014

Jordan E. DeLong, Kaitlin L. Brunick and James E. Cutting, 'Film Through the Human Visual System: Finding Patterns and Limits', in J. C. Kaufman & D. K. Simonton, eds, *Social Science of Cinema*, Oxford

Andrew Goldstone and Ted Underwood, 'The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us', *New Literary History*, vol. 45, no. 3

Holst Katsma, 'Loudness in the Novel', *Stanford Literary Lab*, Pamphlet 7

Sara Klingenstein, Tim Hitchcock and Simon DeDeo, 'The Civilizing Process in London's Old Bailey', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 26

2015

Mark Algee-Hewitt, Ryan Heuser and Franco Moretti, 'On Paragraphs: Scale, Themes and Narrative Form', *Stanford Literary Lab*, Pamphlet 10

James E. Cutting and Ayse Candan, 'Shot Durations, Shot Classes and the Increased Pace of Popular Movies', *Projections*, vol. 9, no. 2

Anne Dewitt, 'Advances in the Visualization of Data: The Network of Genre in the Victorian Periodical Press', *Victorian Periodicals Review*, vol. 48, no. 2

Franco Moretti and Dominique Pestre, 'Bankspeak: The Language of World Bank Reports', *New Left Review* 92, March–April

2016

Mark Algee-Hewitt et al., 'Canon/Archive: Large-Scale Dynamics in the Literary Field', *Stanford Literary Lab*, Pamphlet 11

James F. English, 'Now, Not Now: Counting Time in Contemporary Fiction Studies', *Modern Language Quarterly*, vol. 77, no. 3

Marissa Gemma, 'The Making of Middle American Style: Narrative Talk in the 19th Century Novel', presentation at the Stanford Literary Lab, 15 February

Hoyt Long and Richard Jean So, 'Turbulent Flow: A Computational Model of World Literature', *Modern Language Quarterly*, vol. 77, no. 3

Olivier Morin and Alberto Acerbi, 'Birth of the Cool: A Two-Centuries Decline in Emotional Expression in Anglophone Fiction', *Cognition and Emotion*, vol. 31, no. 8

Ted Underwood and Jordan Sellers, 'The *Longue Durée* of Literary Prestige', *Modern Language Quarterly*, vol. 77, no. 3

2017

Mark Algee-Hewitt, 'Distributed Character: Quantitative Models of the English Stage, 1550–1900', *New Literary History*, vol. 48, no. 4

Marissa Gemma, Frédéric Glorieux and Jean-Gabriel Ganascia, 'Operationalizing the Colloquial Style: Repetition in 19th-Century American Fiction', *Digital Scholarship in the Humanities*, vol. 32, no. 2

Leonardo Impett and Franco Moretti, 'Totentanz: Operationalizing Aby Warburg's *Pathosformeln*', *New Left Review* 107, September–October

Maria Kanatova et al., 'Broken Time, Continued Evolution: Anachronies in Contemporary Films', *Stanford Literary Lab*, Pamphlet 14

Thomas Lansdall-Welfare et al., 'Content Analysis of 150 Years of British Periodicals', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 4

Grace Muzny, Mark Algee-Hewitt and Dan Jurafsky, 'Dialogism in the Novel: A Computational Model of the Dialogic Nature of Narration and Quotations', *Digital Scholarship in the Humanities*, vol. 32, no. 2

2018

Taylor Arnold and Lauren Tilton, 'Distant Viewing: Analysing Large Visual Corpora' [under review]

Jonah Berger and Grant Packard, 'Are Atypical Things More Popular?', *Psychological Science*, vol. 29, no. 7

Jo Guldi, 'Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora', *Journal of Cultural Analytics*, 20 December

Hoyt Long, Anatoly Detwyler and Yuancheng Zhu, 'Self-Repetition and East Asian Literary Modernity, 1900–30', *Journal of Cultural Analytics*, 21 May

Ted Underwood, David Bamman and Sabrina Lee, 'The Transformation of Gender in English-Language Fiction', *Journal of Cultural Analytics*, 13 February