Distance and Depth, Computers and Close Reading

27th May 2020

Cambridge Digital Humanities

I have been struck in the past few years by how computational methods in literary studies have taken off, mostly in the field of digital literary history. We've had Ted Underwood's *Distant Horizons*, Andrew Piper's *Enumerations*, Matt Jockers's *Macroanalysis* and a heap of other works that examine a new, much broader scale of literary history, looking at texts well beyond the traditional canon from which our histories are derived, often with surprising results.

I wanted to know, though, what had happened to the many experiments in digital close reading? I thought of John Burrows's examination of the novels of Jane Austen; Catherine Nicholson's work on breadth or generality and specificity in literary studies; or Tanya E. Clement on Stein's *The Making of Americans* and fractal pattern-making therein. How, I wanted to know, had so-called distant reading become associated in the popular – so far as we can call literary critical studies popular – imagination with large-scale literary history, rather than with close attention to texts?

One of the reasons I find this so interesting is that all distant methods for reading literary history must begin with a close-up focus on the texts themselves. Distant reading at the macro level depends on computational processes that can work at the micro-level and that require verification against human close reading processes. For it makes no sense to believe that a distant reading at scale can be any good unless its processes truly work at the level of the individual texts – and work reliably across a well-known set of extant novels or poems.

It is in this space that much digital humanities work in the literary space is misunderstood. For it is often claimed, when a digital literary studies paper proclaims that it has replicated an existing finding (i.e. the computational approach agrees with existing literary criticism), that these methods tell us nothing new. Yet this is an incremental approach to knowledge that renders computational approaches compatible with extant literary criticism. As Matthew L. Jockers puts it in his introductory instruction book on using the programming language *R* for text analysis: '[a] good deal of computational work is specifically aimed at testing, rejecting, or reconfirming knowledge that we think we already possess'.[1]

Yet to say that such work tells us only what we already know is *to miss the point*. When such work confirms an existing literary critical supposition, it not only adds further confirmatory evidence to that point of view – especially pertinent given that interpretative questions in literary criticism are often pluralistic and allow for divergences of opinion – but also demonstrates that the computational method is, in some way, sound. When such approaches work across multiple texts, we can then have some confidence that they might be accurate when they are scaled to work on texts that we have not read in advance or that do not have existing theories that we might test. When a computational method works on many texts, but then throws an anomaly on another, it can force a reevaluation of the extant theories around that single text. The challenge for digital methods is that this flattening of texts into constellations of relation (literary history) – the knocking down of complex individual works into data points – is done without having read the originals. 'The risk',

1 Matthew L. Jockers, *Text Analysis with R for Students of Literature* (New York, NY: Springer, 2014), pp. vii–viii.

write Claire Lemercier and Claire Zalc, ‘of standardization has always plagued quantitative history’.[2]

I was also interested, though, in what digital approaches might bring that were *new* when applied closely to texts. Computers are, after all, the ultimate instruments for repetitive, brute-force activities. They are really good at conducting *boring*, soul-destroying tasks, such as counting word frequencies. Now, by themselves, such frequencies tell us very little of interest. If you count the number of times the word whale appears in *Moby Dick*, writes Timothy Brennan, you will not learn much more than how many times the word whale appears in *Moby Dick*. However, when combined with analyses of foregrounding, or with comparative frequencies in suitable corpora, I could immediately see uses for such techniques that would tell us more than how many times the word whale occurs in *Moby Dick*.

Furthermore, there are often questions that I have of texts that would be impossibly tedious to answer – but to which I still wish I had an answer. In the case of David Mitchell’s *Cloud Atlas* – the novel on which I focus in *Close Reading with Computers*, for example – I wanted to know whether there were any anachronistic words in the section that purports to be written in 1850. Sadly, though, I lack the patience to lookup tens of thousands of words in various etymological sources by hand. I do, though, have the rather more minimal patience that is required to write a computer program that will do this on my behalf.

Interestingly, in that particular instance, the task and its results throw up an interesting set of theoretical questions about the text itself:

1. Is the text really set in 1850 and what internal evidence should be used to suggest an etymological cut-off date? (there is evidence in the text that 1920 might be a more appropriate and textually consistent first usage date to use as the cutoff)
2. What is an authoritative etymological source when two dictionaries disagree?
3. Are there textual explanations (“the diary was forged!”) that allow such slippages to be explained away?
4. How do we handle the “dirty data” that comes back from such a lookup exercise? (For instance, my script returned the etymological use of “colour” as a term for musical timbre as a false positive for a 1940s word)

In short, the computational approach in no way eradicates or supersedes human interpretation and reading; it simply provides more fodder for interpretative practices.

Of my anachronisms, ‘Lazy-eye’, in particular, is perhaps the most startling find. This term sounds like a pejorative slur for people with amblyopia that would have been coined well before 1960. It’s

2 Claire Lemercier and Claire Zalc, *Quantitative Methods in the Humanities: An Introduction*, trans. by Arthur Goldhammer (Charlottesville, VA: University of Virginia Press, 2019), p. 3.

something we would expect to find in literature from 1850, not in contemporary discourse. However, this is not actually the case and the term was coined far later than we might usually imagine. The fact that we cannot recognise which words are appropriate to a time period brings to the fore a problem that has vexed historical fiction and its study for many years: to what extent is accuracy to the historical record actually important? And if the language is not totally historically accurate, what other markers might signify to a contemporary reader that the work is from the past?

Similarly, counting word frequencies alone is not particularly useful. But what happens when you compare word frequencies – in, say, this 1850 diary segment – with a contemporary magazine corpus? I argue, in the book, that this reveals some traits of how we imagine 19th-century style in the twenty-first century.

For the contemporary writer who wishes to mimic the writing of the past, this all contributes to a type of imagined language that Mitchell has called ‘bygonese’. In particular, when a contemporary novelist attempts to create a plausible pastiche of writing from the past, there is an expectation that things will not be ‘as they were’, but rather, in a knowing fashion, ‘as we imagine they were’. ‘To a degree’, writes Mitchell, ‘the historical novelist must create a sort of dialect – I call it “Bygonese” – which is inaccurate but plausible. Like a coat of antique effect varnish on a pine new dresser, it is both synthetic and the least worst solution’. This comes about because, of course, we know that contemporary fiction is making it up as it goes along even while, simultaneously, we want to believe that the twenty-first-century novel ‘has got it right’.

In order to test this, I took a set of magazine articles from 2004 and pulled out words that occur in *Cloud Atlas*’s Ewing chapter that are not in the magazine. The result is, indeed, a set of terms that are unusual to the modern ear and that sound archaic.

In particular, though, *Cloud Atlas*’s Ewing chapter falls back on offensive racial addresses in order to achieve its historical style and its critical focus on the legacies of colonialism. For instance, Mitchell’s text gives us: Blackamoor, blackfella, darkies, harridan, womenfolk, bedlamite, mulatto, quadroon, and mixedblood. Specifically, colonial terms of racist abuse occur in the Ewing section of *Cloud Atlas* at a far-higher frequency than in a broader contemporary corpus. These are used, in the text, I should stress, for purposes of critique – not to condone such language and its imperial origins.

There’s a great deal of debate about whether objects in texts are good markers of their genre. For instance, Ted Underwood has recently pointed out that science fiction novels are more clearly defined by their use of the adjectives of scale, rather than spaceships and so forth (although they do often have the latter). In *Cloud Atlas*, though, racial epithets serve to build an imperial, Empire-based racist charge in the language that contributes to our belief that this writing could really have come from 1850. This is strengthened through the affiliation with outmoded colonial-era notions of

‘tropical medicine’ in this novel, in which the white man may fall prey to the diseases of the warmer climes.

All of which is to say that *Cloud Atlas* makes for an excellent case study of how contemporary writers build an aesthetic of how we wrote in the past. But the type of close attention that we require in order to make this work is made easier through the add-on tool of computational counting and approaches. In the book, I keep my computation relatively simple. I do not make any claims to the sophisticated predictive modelling of Underwood or Piper, but instead simply wondered how I could make the computer do the heavy lifting, to drive empirical questions