Justin Tonra,
NUI Galway

Poetry by the Book, Poetry by Numbers[1]

1

The mass digitisation of our literary heritage has resulted in both possibilities and problems for the literary scholar. With the availability of large-scale corpora of literary texts comes the implicit perception that digital surrogates yield the same information as the physical object from which they were produced. But this is not the case. As scholars, we stand to lose a great deal, including accuracy and credibility, by turning our backs on the material book. At the same time, however, increased access to masses of literary data enables scholars to make computer-assisted queries that cannot be made by an individual human brain. Computers can process huge quantities of text and, with scholarly guidance and interpretation, provide fresh insights into our understanding of literary history.

This talk describes examples of the apparently contradictory approaches to literary study represented by the computer and the book, and suggests that they are more similar than they appear. Both are at heart inspired by a philological imperative to preserve our cultural heritage and provide a means for its investigation.

2

Literary criticism, and academia generally, has often marshalled the serendipitous discovery in the service of constructing an argument or a critical judgment. Such serendipity can take material

[1] Part of the title of this talk is borrowed from Andrew Stauffer, who teaches a course called 'Poetry By the Book.'

or cognitive form (the manuscript found in the attic; the eureka moment), and provide the raw materials for analysis and conjecture. In literary criticism, arguments are generally based upon evidence gleaned from close reading in support of a hypothesis, but can the literary text yield evidence that is not immediately discernable to the human eye for a similar interpretive purpose? Computers have produced such data in the form of concordances and word-frequency lists for about half a century, but with the greater availability of tools for these purposes, the *scholarly* value of these tasks is diminished. How can we use the discoveries of such computing processes for literary criticism, for articulating the meaning in a text? Can literary criticism and computing enrich one another? Is algorithmic criticism—criticism derived from algorithmic manipulation of text (Ramsay 2)—possible?

Interpreting the relationship of textual part to textual whole and thence making critical inferences is a relatively unchallenged literary orthodoxy. It has successfully precluded challenge on the basis that it is a faithful and honest engagement with the text; however, the selectivity inherent in such criticism often amounts to a radical reworking of the text. Stephen Ramsay gives two examples: first, Chinua Achebe's focus on representations of Africans in *Heart of Darkness* 'reinscribes' the novel, he suggests, to "make the text tell a different story" (42). And second, Jerome McGann and Lisa Samuels' critical investigation of what is gained, in interpretive terms, by *literally* reading a poem from back to front. These examples query whether one form of textual deformation is more valid than another, by comparing a traditional method with one that is apparently radical and counter-intuitive. The essential point, however, is that the act of deforming the text is inherent in all types of literary criticism. In digital text analysis this fact is explicitly acknowledged.

Algorithmic criticism involves a reorientation of the critical paradigm as well as a deformation of the text. Instead of advocating a close readerly engagement with the text, it dispenses with reading altogether. This is one of the reasons why it provokes discomfort: it interrupts the sacrosanct relationship between reader and text, and arouses a range of culturally and professionally inherited anxieties that we have about reading. Literary scholarship, as we know, is not about reading everything, but about being able to find one's bearings within literature as a system. Distant reading (as theorised by Franco Moretti) is a step towards making empirical facts about a text the basis for literary criticism. This does not always begin *without* an interpretive instinct (as Stanley Fish has recently suggested), but it provides a means for thinking about a critical instinct and interrogating a text (or texts) to see whether it will yield evidence to develop that instinct into an argument.

But is there is a paradox in using digital text analysis to explore cultural and historical contexts? The Humanities tolerate computer-aided textual analysis in disciplines such as authorship attribution because it adjudicates in a question that usually results in a positive or negative outcome. Where a computer is involved in a question that moves from this simple binary to the hermeneutics that is more common in the Humanities, suspicion soon follows. However, algorithmic criticism can bridge the divide created by this suspicion. It can use empirical evidence from algorithmic processes to support hypotheses born of hermeneutic inquiry, and may also provoke fresh interpretations of raw data. Or, as Ramsay asks, "Why can't data function within the same interpretative regime —— and fulfil the same hermeneutical functions —— as text[?]"

3

Let's take an example. I have used digital text analysis in order to study coincidence, influence, and imitation between the Romantic Orientalist writings of Lord Byron and Thomas Moore [SLIDE]. I thought fresh empirical insight into this topic desirable because of the difficulty of writing about these issues in a way that it not speculative or nebulous. Speculative assertions based upon close reading seemed quite limiting to me in this case: it demands firstly a heuristic, not a hermeneutic approach. Is Romantic Orientalism defined by a limited vocabulary which leads to similarities and coincidences between its practitioners? Does writing within a specific genre impose *formal* or *semantic* constraints upon the author? And, to interrogate Romanticism on its own terms, are there ways in which the skilled practitioner always transcends the limitations of genre to approach a greater poetic Truth? To my mind, algorithmic criticism enables a means of thinking differently about these issues than can be achieved through close reading alone.

While analysing a corpus of five long poems by Byron and four by Moore from the period 1813 to 1817, I noticed a consistently high relative frequency of the word *eyes* that aligned it closely with Moore's poems [GRAPH]. However, graphing the relative frequency of *eye* highlighted an equally prominent association of the singular noun form with Byron's writing [GRAPH]. Pursuing this pattern further revealed a general trend across the corpus[2]. When plotting a sample of corporeal and corporeal-related words used by both authors, two notable facts are obvious: there are very distinct and separate clusters of Byron and Moore texts, and consistent association

[2] Besides *eye(s)*, noun pairs which exhibit this associative tendency include *cheek(s) heart(s) lip(s) tear(s) voice(s)*.

of singular nouns with Byron and plural nouns with Moore [GRAPH]. [ScatterPlot is a tool that creates a graph of documents and terms, spaced by their variation from one another. Terms appear in closer proximity to documents in which they are more frequent, and the larger the circle, the higher the term's frequency].

Though they are not the only senses of the word[3], Byron uses *eye* for metaphoric or metonymic effect [SLIDE]. Indeed, though the singular form rarely appears in his texts, Moore utilises it for the same purpose. When the plural form is used, as it is more frequently by Moore, it refers more often to the organic, corporeal meaning than does the singular form [SLIDE]. The question that this has led me to ask is whether the singular noun form lends itself more naturally to *poetic diction* than the plural? I have not yet completed a fuller survey of the noun forms in the corpus, but the explicit division of the forms between the two poets is remarkable.

But my more general query today is this: can we be comfortable with this computer-generated data as valid evidence for interpretation? Here is a statement of the kind that appears in every academic paper: 'It is clear that Byron effectively invokes the metaphysical through the physical (*breast face hand*), and since this effect is more limited by Moore's more literal favouring of *hearts* and *smiles*, it is no surprise that his tales command the majority of *soul* and *spirit* in both singular and plural forms. In this context, these words are not only a token of Moore's theism (as opposed to Byron's atheism), but of his specific poetic means of engaging the metaphysical.' This is my interpretation of evidence from the texts, but it could not have been made on the basis of reading the texts alone (not matter how attentively). Without the computer's ability to quantify

[3] The singular and plural forms of the word each occur more than 150 times in the corpus, so it is difficult to state that a particular use or sense of either word is representative or dominant.

vocabulary and visualise its associative relationships—"to assist the critic in the unfolding of interpretative possibilities," as Stephen Ramsay has it (8), the basis for the assertion would have remained dormant and undiscovered within the texts.

This analysis, based on nine long poems, is comparatively small-scale. Ramsay works with a smaller dataset while investigating different vocabulary patterns of Virginia Woolf's *The Waves*. But others have worked with far larger datasets to explore our understanding of *longue durée* literary history. Matt Jockers [BOOK] has used computers to see "to what extent a nineteenth-century Irish voice may be heard (or measured) amid a cacophony of 1,030 British and Irish novels" (112). Examining a number of different linguistic and stylistic features in the corpus, Jockers identifies evidence for a familiar postcolonial reading of Irish (literary) history: that British novels more frequently contain words denoting 'determinacy' and 'confidence,'[4] and first-person pronouns. Meanwhile, Irish novels more frequently use words characteristic of 'imprecision' and 'indeterminacy,'[5] and underutilise first-person words. Jockers, whose primary interest is digital methods, does not shape this data into a formal argument, but suggests that it could be adopted in support of differing conceptions of Irish literary history: one, that Irish novelists were playing catch-up with their British counterparts (writing what Terry Eagleton has called an "imperfect realism," set against the supposedly paradigmatic realism of nineteenth-century British novel), or, two, that other factors (social, cultural, historical?) resulted in a unique Irish fictional consciousness that manifested itself in uncertainty and a lack of self-reflection.

[4] *always should never sure can cannot*

[5] *near soon some less more much*

Neither of these literary-historical narratives is unfamiliar. What is different is the (empirical) means by which the evidence in support of these narratives is reached.

4

Textual corpora like Jockers', once the preserve of linguists, are increasingly being used as a source for literary scholarship. Not only does this have consequences for established canons and reading patterns, but it can also challenge other scholarly assumptions about periodicity. Given the ease of access to Eighteenth-Century Collections Online, is the eighteenth-century scholar now obliged to engage with the thirty-three million pages of books from the period in ECCO? Perhaps not, but in the face of such abundance, he or she may sense the inadequacy of some received critical assumptions about the period that are founded upon observations about a relatively small set of texts. ECCO is useful insofar as it grants access to facsimiles of eighteenth-century books for scholars not fortunate enough to have ready access to a well-stocked research library (those poor souls). In its 200,000 volumes, it includes the majority of books that were published in the UK in the eighteenth century (as well as thousands of titles from other countries). Its coverage is extremely wide—your library doesn't have a copy of Pope's *Dunciad* of 1728? Luckily for you, ECCO does (it has five). But ECCO is not just useful for enabling access to individual copies. It also calls the scholar to consider the mass, the totality, the entire literature of the eighteenth century—but it also hampers certain efforts to do so. It has a Google-style search interface that is oriented towards finding text in individual copies, rather than revealing broader patterns. It is not possible from ECCO's proprietary search to examine the comparative signification of *eye* and *eyes* across the entire corpus. However, the Text Creation Partnership has made ECCO's texts freely available for download in a variety of formats, thus

providing the basis for customised searching and algorithmic manipulation of the corpus. Many of the next steps in this process are technical, and the means of achieving them part of another discussion, but the importance of freely available, open-source literary texts cannot be underestimated.

Researching literary history with a corpus of 200,000 texts does mean certain sacrifices, however. And one sacrifice is accuracy: just as the manual transmission of texts through printing introduces errors, so does the digitisation of texts [SLIDE]. And while it is possible to research and uncover the textual history of a single work, the task of assessing the accuracy of 200,000 texts is overwhelming. And this is why the preservation of our material literary heritage demands as much attention as its digitisation. Because *both* enable the recovery of literary history through opposing methodological poles: computing 1,030 novels [JOCKERS' BOOK]; examining physical copies of one poem [DVM's BOOK on the 1728 *Dunciad*].

5

To return to Pope's 1728 *Dunciad*, ECCO contains surrogates of five copies printed that year. Four of those come from different states of the first three editions[SLIDE], but they fail to reveal the complex textual and transmission history of the *Dunciad* in that year. Why? Because they are individual copies that claim to represent a full edition, but in the eighteenth century it was common for copies apparently from the same edition to exhibit variation resulting from changes made in the course of printing. For example, the first impression of the third edition (which contains a different title-page ornament to the second impression, as well as a host of textual variants) is not available in ECCO. Nor is the variant state of the second edition, where the title-

page misspells ‘DUBLIN’ as ‘DUDLIN.’ The duodecimo and octavo impressions of the first edition were printed concurrently in parts, but the bibliographical task of determining the precedence of the duodecimo was partly achieved by examining material and paratextual forms of evidence (such as page and type page dimensions) that are obscured in the digital copy. These are pedantic points, but happily pedantry is a virtue in bibliography, and recovering the early transmission history of one of the century’s major poems is no trivial matter. This matter is resolved (and the entries in ECCO point to Vander Meulen’s work) only because physical copies of these books are extant (Vander Meulen consulted sixty-seven). Future research of this kind will simply not be possible by examining digital surrogates alone. Digitisation of our literary heritage is doubtless a positive move, but it is not a *substitute* for its physical preservation. Jerome McGann struck an advisory note in *Critical Inquiry* earlier this year: “Consider for a moment two of philology’s most elementary methodological rules: first, in researching a problem, always examine the original materials *in situ*; second, consult the primary materials at least twice, once at the beginning and again at the end of your work. Surrogates, digital or otherwise, may serve the work at some point, but they cannot substitute for those first-hand visits.” If our cultural inheritance has a future (and this reaches beyond literary studies to all areas of the Humanities), it must ultimately be in the union of the material and the digital [SLIDE].

20 May 2013
NUI Galway.