Renaissance and Reformation
Renaissance et Réforme

Fieldwork in the Sonnet: Milton, Donne, and Critical Orthodoxy

Michael Ullyot

Volume 44, numéro 3, été 2021

Digital Approaches to John Milton
Approches numériques de l'oeuvre de John Milton

URI : https://id.erudit.org/iderudit/1085821ar
DOI : https://doi.org/10.33137/rr.v44i3.37989

Aller au sommaire du numéro

Éditeur(s)

Iter Press

ISSN

0034-429X (imprimé)
2293-7374 (numérique)

Découvrir la revue

Citer cet article

Ullyot, M. (2021). Fieldwork in the Sonnet: Milton, Donne, and Critical Orthodoxy. *Renaissance and Reformation / Renaissance et Réforme*, *44*(3), 25–43. https://doi.org/10.33137/rr.v44i3.37989

Résumé de l'article

Dans cet article, Michael Ullyot explore les possibilités, pour la lecture littéraire, qu'engendre la disponibilité d'une vaste base de données de sonnets. Ullyot suggère qu'une bibliothèque élargie de textes et d'outils d'analyse nous aiderait à lire le sonnet de manière moins linéaire et plus « évolutive ». Bien qu'il ne soit pas encore complètement développé, le prototype d'Ullyot a déjà fourni des résultats concluants en se concentrant sur le corpus, certes restreint, des sonnets anglais de Milton. Si les techniques d'analyse textuelle peuvent confirmer des intuitions critiques sur les sonnets de Milton, comme le suggère Ullyot, nous pourrons alors davantage avoir confiance en ces outils et techniques lorsque, plus tard, ceux-ci seront étendus à des ensembles de textes plus importants. Tout en faisant preuve d'un pragmatisme opportun quant aux défis à venir, l'article d'Ullyot laisse entrevoir toutes les retombées positives qui pourraient résulter de la plus grande anthologie de sonnets au monde.

Fieldwork in the Sonnet: Milton, Donne, and Critical Orthodoxy[1]

MICHAEL ULLYOT
University of Calgary

In this article, Michael Ullyot explores the possible implications for literary reading of a vast textual database of sonnets. Ullyot argues that a growing library of texts and tools will help us to read the sonnet in less linear and more "scalable" ways. Although not fully developed yet, Ullyot's prototype already has produced useful results by focusing on Milton's admittedly small corpus of English sonnets. Ullyot suggests that if textual analysis techniques can confirm critical insights about Milton's sonnets, then more confidence can be placed in these tools and techniques when, later, they are scaled up to larger bodies of texts. While displaying a healthy pragmatic realism about the challenges ahead, Ullyot's article tantalizingly suggests the scholarly advantages of building the world's largest sonnet anthology.

Dans cet article, Michael Ullyot explore les possibilités, pour la lecture littéraire, qu'engendre la disponibilité d'une vaste base de données de sonnets. Ullyot suggère qu'une bibliothèque élargie de textes et d'outils d'analyse nous aiderait à lire le sonnet de manière moins linéaire et plus « évolutive ». Bien qu'il ne soit pas encore complètement développé, le prototype d'Ullyot a déjà fourni des résultats concluants en se concentrant sur le corpus, certes restreint, des sonnets anglais de Milton. Si les techniques d'analyse textuelle peuvent confirmer des intuitions critiques sur les sonnets de Milton, comme le suggère Ullyot, nous pourrons alors davantage avoir confiance en ces outils et techniques lorsque, plus tard, ceux-ci seront étendus à des ensembles de textes plus importants. Tout en faisant preuve d'un pragmatisme opportun quant aux défis à venir, l'article d'Ullyot laisse entrevoir toutes les retombées positives qui pourraient résulter de la plus grande anthologie de sonnets au monde.

> … and, when a damp
> Fell round the path of Milton, in his hand
> The Thing became a trumpet; whence he blew
> Soul-animating strains—alas, too few!
>
> William Wordsworth: "Scorn not the Sonnet; Critic, you have frowned"

The way we read now

In this journal's fiftieth anniversary issue of 2014, I speculated about future methods of critiquing early modern literature. Progress would come from

1. Dedicated to the memory of Joshua James Harkema (1984–2019).

https://doi.org/10.33137/rr.v44i3.37989

scholars using text-analysis technologies to access archival materials, I wrote, particularly the Early English Books Online-Text Creation Partnership (EEBO-TCP) corpus.[2] If we could search such a corpus for any domain of inquiry—from topics to, say, poetic forms—then our arguments about that domain would be based on more evidence. Those arguments would be more wide-ranging than narrow, more definitive than provisional.

That future has not yet arrived. Critics today still make provisional arguments on narrow subjects. That is partly because humanists resist generalizations as inherently untrustworthy, even those founded on wide-ranging evidence. "Skepticism about generalization might even have come to constitute the most basic mission of the humanities," writes Caroline Levine.[3] We value what Jan Parker calls "a hermeneutic of concentration": "one of deriving interpretative narratives, often plural, avowedly partial, from singular, particular events."[4] Literary critics focus on a limited set of singular objects whose particularities evidence arguments that do not range far beyond them. Even if we could do nuanced searches of a corpus with tens of thousands of texts, we probably wouldn't.

In her 2019 omnibus review essay "Recent Studies in the English Renaissance," rightly described as "the state of the union," Catherine Bates mentions "current methodological practices within the discipline, the most dominant of which remains historicism."[5] She makes a witty comparison of Stephen Greenblatt to Martin Luther, charismatic leaders who both ushered in transformative new orthodoxies. She briefly gestures toward "Rapid technological developments in the *accessing* and *disseminating* of information [that have] no doubt helped it on its way. The rest, as they say, is historicism."[6]

2. Michael Ullyot, "Augmented Criticism, Extensible Archives, and the Progress of Renaissance Studies," *Renaissance and Reformation / Renaissance et Réforme* 37.4 (2014): 179–93, dx.doi.org/10.33137/rr.v37i4.22646.

3. Caroline Levine, "Model Thinking: Generalization, Political Form, and the Common Good," *New Literary History* 48.4 (2017): 633–53, 633, dx.doi.org/10.1353/nlh.2017.0033.

4. She continues, "With digital data, however, there is, rather, a sense of dizzying plurality, seemingly stretching almost to infinity." Jan Parker, "Speaking Out in a Digital World: Humanities Values, Humanities Processes," in *Humanities in the Twenty-First Century: Beyond Utility and Markets*, ed. Eleonora Belfiore and Anna Upchurch (New York: Palgrave Macmillan, 2013), 44–62, 51–52, dx.doi.org/10.1057/9781137361356_3.

5. Catherine Bates, "Recent Studies in the English Renaissance," *Studies in English Literature 1500–1900* 59.1 (2019): 203–41, 213, dx.doi.org/10.1353/sel.2019.0009.

6. Bates, 213; my italics.

Bates's claim that technologies are instrumental to the dissemination of information is accurate. It is easier for critics of early modern literature to access primary materials when they are digitized. But "accessing and disseminating" technologies only deliver texts that we read, annotate, and cite as if they were in print. Our methods progress far more incrementally than our knowledge does. The discipline values print over other media, so we use word-processing programs to prepare manuscripts for print.[7] We make knowledge by reading texts, usually in print. We annotate margins. We gather exemplary segments of textual evidence to make cumulative arguments. We read what past critics, using the same techniques, have written about those texts and other texts. In short, our ways of accessing texts and disseminating ideas about them—the habits and the *habitus* of criticism—would not be alien to their early modern authors. Bates's technologies are digital versions of the book-wheel.

I say this without hand-wringing. In 2014 I overestimated the pace and necessity of methodological progress, and will not commit that error here. This article is not a jeremiad against criticism's time-proven methods, but endorses their value by extending their scope: critics' technologies of access should extend past the level of documents to the level of words. The ease with which we search library catalogues or the *Oxford English Dictionary* database, for example, should extend to the contents of our documents. Critics tend to neglect text technologies that go beyond text- or entry-level access, that deform and quantify and manipulate the linear arrangement of words.[8] I suggest that we address ourselves to their words, when and if they can accommodate our queries.

Using technologies on the words of our texts has twofold implications: for the direction of our attention (what to read) and the methods of our attention (how to read). Before I address them in turn, let me address a cognitive bias of both digital and traditional critical methods: the law of the instrument, summarized in the glib axiom "To a hammer, every problem looks like a nail." Enthusiasts for digital tools are particularly susceptible to this bias. In 2020, the year after Bates's review essay, Ryan Netzley's own *SEL* survey of sixty-three monographs, editions, and collections turned to David Currell and Islam Issa's

7. The term "print" in my usage encompasses digital surrogates that replicate print media, like ebooks and electronic journals.

8. L. Samuels and J. J. McGann, "Deformance and Interpretation," *New Literary History* 30.1 (1999): 25–56, dx.doi.org/10.1353/nlh.1999.0010.

Digital Milton, to which many articles in this special issue are indebted.[9] Netzley says that too many digital tools turn "interpretation into a handmaiden" for their self-justifying arguments: "the literary problem doesn't seem to exist before the apparatus appears to solve it; rather, the literary problem conforms to the apparatus."[10] It's hard for an omnibus review to do justice to every essay in every collection without betraying a few preconceptions, yet Netzley's claim gives me pause—because it diagnoses a tendency of digital criticism. The law of the instrument explains why so much ink is spilled on the attribution question in Shakespeare studies.[11] It's not because scholars are desperate to learn what stylometry reveals about where Shakespeare ends and Middleton or Ford begins; it's because stylometry offers a definitive answer to a question—even if few people were raising that question.

Before you agree too vigorously with that diagnosis, gentle reader, heal thyself. The law of the instrument is more pervasive and subtle than you recognize. The instrument is our own cognition, imposing seemingly natural critical methods like linear reading and heuristic intuitions. Just as we fail to notice that written language or the codex or corrective lenses are technologies, we fail to recognize that our critical methods are not necessary conditions of interpretation. When we could use other methods, ours are not exclusively necessary, and far from sufficient. Humans have limits: to our memories, and to our reading procedures. We can only read and retain so much, by moving sequentially through texts, accumulating evidence to make arguments. Susan Hockey has described this as "the somewhat serendipitous noting of interesting

9. David Currell and Islam Issa, eds., *Digital Milton* (Basingstoke: Palgrave Macmillan, 2018), dx.doi.org/10.1007/978-3-319-90478-8.

10. Ryan Netzley, "Recent Studies in the English Renaissance," *Studies in English Literature 1500–1900* 60.1 (2020): 153–97, 180, dx.doi.org/10.1353/sel.2020.0007.

11. The most noteworthy studies include Jonathan Hope, *The Authorship of Shakespeare's Plays* (Cambridge: Cambridge University Press, 1994), dx.doi.org/10.1017/CBO9780511518942; Brian Vickers, *"Counterfeiting" Shakespeare: Evidence, Authorship, and John Ford's* Funerall Elegye (Cambridge: Cambridge University Press, 2002), dx.doi.org/10.1017/CBO9780511484049; D. H. Craig and Arthur F. Kinney, *Shakespeare, Computers, and the Mystery of Authorship* (Cambridge: Cambridge University Press, 2009), dx.doi.org/10.1017/CBO9780511605437; Mireille Ravassat and Jonathan Culpeper, eds., *Stylistics and Shakespeare's Language: Transdisciplinary Approaches* (London: Continuum, 2011); and MacDonald P. Jackson, *Determining the Shakespeare Canon* (Oxford: Oxford University Press, 2014), dx.doi.org/10.1093/acprof:oso/9780198704416.001.0001.

features."[12] We do it because of what Jonathan Hope and Michael Witmore call our "linear reading and the powerful directionality of human attention."[13] But our attention is no longer a necessary constraint to our interpretations, and the era is past when linear readings have exclusive dominion over critical methods.

We can de-anthropomorphize literary criticism without dehumanizing it, however, by retaining what's valuable about human cognition: its grasp of nuance, or of verbal connotations, or of narrative shape. I mention the lattermost because Hope and Witmore's argument—that Shakespeare's *Othello* has linguistic features, discernible only to a computer—aligns the play with comedies rather than with his tragedies. Despite this interesting feature, human critics are not recategorizing the play as a comedy; when assigning genre we address it at the document level, not the sentence or word level. The story's qualities matter more than the computer's quantifications of its language, as least for the genre question. The trouble with such methods, valid as they are for determining genre, is their narrowness. Human practices of linear reading permit us insufficient time to read everything, so we rely on the tyranny of custom—or more precisely, of canons. Andrew Piper has shown that in 2015 the 1 percent most studied authors were subjects of a fifth of the MLA International Bibliography's 6,252 articles or book chapters on literary studies, while the top 20 percent of authors accounted for just under 60 percent.[14] Scholars trust teachers, editors, and other critics to guide our attention. The result is more research on canonical authors, reinforced by an interpretive community also familiar with those authors.

For scholars of early modern literature, new historicism has opened the archive. Yet the list of primary authors who earn critical arguments remains narrowly canonical. Wyatt and Surrey, Sidney and Spenser, Shakespeare and Donne all deserve and repay close readings. But every reading of them is a

12. Susan Hockey, *Electronic Texts in the Humanities* (Oxford: Oxford University Press, 2000), dx.doi.org/10.1093/acprof:oso/9780198711940.001.0001.

13. Jonathan Hope and Michael Witmore, "The Hundredth Psalm to the Tune of 'Green Sleeves': Digital Approaches to Shakespeare's Language of Genre," *Shakespeare Quarterly* 61.3 (2010): 357–90.

14. Andrew Piper, *Enumerations: Data and Literary Study* (Chicago: University of Chicago Press, 2018), 183.

choice to exclude others.[15] Are we confident that those exclusions are deserved? Put it another way: are we allocating critical attention on the right grounds?

Whether or not we recognize it, there is an arbitrary selectivity to canons. They are the evidence that happens to be at hand, not necessarily the most suitable evidence for a particular argument. Shakespeare's plays and poems, for instance, comprise about 865,185 words—a mere fraction of 1 percent of the billion words printed in English before 1700. Print them all in a four-hundred-page book, and Shakespeare's contribution would be just over a third of a page. Most scholars could hope in their careers to read perhaps four of those pages. Shakespeare is rich territory for many arguments, obviously, but even his "personal authority" constricts a critic's breadth.[16] That is one advantage to using machines to extend our queries, as I say: they direct our attention to evidence that is not arbitrarily selective.

A second means of judiciously de-anthropomorphizing criticism is to rely less heavily on heuristics, on loose or precise intuitions and definitions that accrete through our training, teaching, and conversations. The example that this article concerns is the English sonnet, which adapts Italian features of form and genre to a cluster of subjects. Think, for a moment, of a sonnet. Not a particular sonnet, with its particular words and themes, with gilded monuments or imagined corners, but a notional sonnet. What does it look like? How does it operate? And where does it originate? Perhaps it is from Stephen Booth, or the *Norton Anthology of Poetry*. Mine comes from my memory of a photocopied handout with Wyatt and Surrey in one column, Sidney and Spenser in the other, dividing Petrarchan octaves from Shakespearean quatrains. It comes from reading and teaching and memorizing canonical examples, the same ones I cite to teach students the conventions of the form. Experience informs epistemology. My exposure to poems called sonnets informs some orthodoxies: that they are primarily a form, secondarily a genre; that they are usually fourteen-line poems in Petrarchan or Spenserian or Shakespearean rhyme schemes; that they typically feature a first-person speaker in "dialectical

15. Franco Moretti, "The Slaughterhouse of Literature," *Modern Language Quarterly* 61.1 (2000): 207–27, dx.doi.org/10.1215/00267929-61-1-207.

16. Andrew Piper, "Think Small: On Literary Modeling," *Publications of the Modern Language Association of America* 132.3 (2017): 651–58, 654, dx.doi.org/10.1632/pmla.2017.132.3.651.

self-confrontation."[17] When my orthodoxies align with other critics', through exchange and negotiation, they are valid.

What choice do human critics have but to be anthropomorphic, to rely on heuristics for our sense-making? The alternative, to locate and read every sonnet ever written, is unfathomable. We can read just one thing at a time, or hold a certain amount of evidence in our head, and therefore we ought to. We may have an interest in Milton, but also in the way he structures prose arguments; we may want to compare their rhetoric with Donne's sermons, or Lancelot Andrewes's. In each case we rely on provisional grasps of authors and texts, forms and genres. Moreover, as Levine and Parker have asserted, we relish particularities even as we leverage them to connect texts in synthetic arguments. The germ of this article, for instance, was Wordsworth's sonnet about sonnets from which I've taken my epigraph. I read it in a Norton anthology called *The Making of a Sonnet*, whose editors compiled an opening section ("The Sonnet in the Mirror") to induce some readerly self-consciousness through the poets' own self-consciousness.[18] Wordsworth led me to Milton's "alas, too few" sonnets (a mere twenty-four), which have an uneasy relationship to more familiar contemporary sonnets of Shakespeare and Donne. The Norton anthology reinforces my orthodoxies about sonnets, while challenging them with unorthodox specimens like Ralph Waldo Emerson's "Woods: A Prose Sonnet." Despite its defiant title, Emerson's sonnet seems deliberately to break every known convention—or known to me, at least.

But I ought to be able to think beyond myself, beyond my particular orthodoxies assembled from limited comparative readings—not by doing more of the same, but by starting from first principles:

1. There is a category of poems that poets and critics call sonnets.
2. There is enough consensus among anthology editors, authoritative critics, and other expert readers to designate some poems as indisputable sonnets.

17. Paul Oppenheimer, "The Origin of the Sonnet," *Comparative Literature* 34.4 (Autumn 1982): 289–304, 299, dx.doi.org/10.2307/1771151.

18. Edward Hirsch and Eavan Boland, eds. *The Making of a Sonnet: A Norton Anthology* (New York; London: W. W. Norton, 2008). The meta-apotheosis is Peter Dickinson's "sonnet on the sonnet on the sonnet" (70).

3. Some sonnets are more conventional than others, insofar as they share features, especially formal features, with many other indisputable sonnets.

Following the lattermost principle, it is possible to compare indisputable sonnets to one another in order to derive a list of their common features: number of lines (usually fourteen); meter (usually pentameter); rhyme (a list of schemes and subdivisions); forms of address (usually first-person); modes of address (usually confessional or observational); and so on. This list grows more uncertain the further it strays from formal to generic features—or, put another way, the more it moves from quantitative to qualitative features. Departure from the conventional number of lines is the readiest measure of a sonnet's unconventionality. (The Norton anthology's editors cleverly title the closing section with outliers: "The Sonnet Goes to Different Lengths.") Yet the number of lines is only one feature among many.

Critics use orthodoxies to answer the question, "Is this a sonnet or not?" This article extols the benefits of quantifying those orthodoxies. If you say that both rhyme and vocabulary make a sonnet, for instance, I will ask to what degree it is rhyme, to what degree it is vocabulary, and to what degree it is other features. I will build a scoring rubric of such features, clustered into formal (rhyme, meter, and length, say) and generic categories (vocabulary, tone, and mode of address, say). This rubric turns the qualities of indisputable sonnets into a quantitative model more accurate than is dreamt of in your critical orthodoxies.

Before I elaborate on this ambition, though, I will focus on the subject of this special issue: the sonnets of John Milton. Databases afford the ability to compare entries in various categories, so I will compare Milton's words to those of other sonneteers in the English tradition—in order to concretize the claims of authoritative critics that Milton's sonnets are atypical. Then I will return to the quantitative model of the sonnet in English, and address its two benefits: metrics of known sonnets, and discovery of unknown sonnets. The latter would achieve the project of surpassing the selectivity of canonical sonnets by formalizing (quantifying) their features into a model suitable for a computer to search large text corpora for unknown sonnets. Anthologies and editions are like the natural history museum, its drawers full of classified specimens; corpora are like the rainforest, brimming with undiscovered species. The time is nigh for sonnet-readers to do more field research.

The diction of Milton's sonnets

In 2018 and 2019, my students in two undergraduate courses (in digital humanities and in seventeenth-century literature) built a database of English-language sonnets. The rationale was that we needed a way to isolate sonnets from other poetic forms in text corpora so that we could compare features only of this subtype of poetry. Sonnets in other corpora like Chadwyck-Healey's *Literature Online* or in individual titles of Project Gutenberg were undifferentiated from adjacent poems and paratexts. We began with sonnets in regularized forms and spellings so we could add text-level metadata like author, period, and source, and token-level metadata like lemmas and rhyme schemes.

In the first class (2018), we built a proof-of-concept SQL database designed by Josh Harkema, a particularly capable former computer scientist to whose memory this article is dedicated. It stored sonnets in TEI-XML, or as JSON objects, and maintained a Python class for data reuse via the RESTful API. (That database has since been absorbed into a temporally broader dataset unavailable for distribution under Canadian copyright law.) The text-level metadata was quite light: author, time period (in fifty-year increments), copyright status, and some back-end details (user, date, etc.) for confirmations and attributions.

The first database included just 445 sonnets transcribed from the aforementioned Norton anthology, Hirsch and Boland's *The Making of a Sonnet*. This anthology offered a selective, sequential history of this poetic type that would allow us to quantify features of (at least) this subset: including formal features (lines, rhyme schemes, syllables, clauses, and sentences) and thematic features, mostly diction-focused (lemma and token frequencies). Such features of the subset would define the model of what a sonnet is, at least in the minds of one anthology's editors; my ambition was that this model would be quantifiable, and thus extensible beyond the subset of these 445 sonnets to un-anthologized sonnets. That remains my governing ambition: a tagged set of known quantities that constitutes a training set for machine-learning algorithms to distinguish sonnets from non-sonnets, insofar as those potential sonnets are quantifiably similar to known sonnets.

There you see the problem inherent in any classification project: you must begin with arbitrary criteria in order to surpass them. In the case of the sonnet, you must begin with form (fourteen-line poems with quatrains) in order to collect as many indisputable sonnets as possible before you extend

your attention to diction and other internal criteria in order to surpass formal criteria. (The division is more complex than I suggest, because formal features like rhyme and meter influence diction.)

To build the second database (2019), my students transcribed the texts of 1,895 early English sonnets by approximately twenty-seven English authors from the sixteenth and seventeenth centuries.[19] They regularized each sonnet with line breaks, punctuation, and spelling to make it look exactly as it appeared in the various edited collections that I selected as our sources: modernized editions, published by Penguin and Oxford, to allow for intertextual comparisons that avoid the irregularities of early modern spelling. We then used a system of double-blind mutual confirmation to ensure that any transcription errors were corrected before the sonnets were accepted into the database.

A few features of this expanded database bear further explanation. Why, for instance, did we use modernized editions whose editorial standards might be inconsistent? My reasoning was that we should use recent editions that are acceptable to scholars in the interpretive community who would gladly accept citations of John Donne's sonnets from A. J. Smith's 1977 Penguin Classics edition—just as they would accept those from (say) C. A. Patrides's 1985 or Ilona Bell's 2006 editions.[20] We can argue about the relative merits of different editions, but my hope was that orthographical or other differences would be mitigated by the vast majority of editions being from the mid- to late twentieth century. In other words, so long as most of their tokens were modernized consistently, we could compare the diction of two authors: say, Thomas Campion (edited 2001) and Thomas Wyatt (edited 1978).

We explored the contents of Project Gutenberg and of Chadwyck-Healey's *Literature Online*, whose sonnets consist mostly of Victorian editions that have fallen out of copyright. Those sources presented a few problems that make those editions unsuitable for scholarly citation: they were edited with

19. The authors are William Alabaster, Philip Ayres, Barnabe Barnes, Richard Barnfield, Nicholas Breton, Thomas Campion, Henry Constable, Samuel Daniel, Michael Drayton, John Davies, John Donne, William Drummond, Fulke Greville, George Herbert, Henry Howard Earl of Surrey, Ben Jonson, Thomas Lodge, Thomas Middleton, John Milton, William Percy, Walter Ralegh, William Shakespeare, Philip Sidney, William Smith, and Thomas Wyatt. The number of authors is approximate because the database included three unattributed sonnets from *The Phoenix Nest* (1593) anthology.

20. A. J. Smith, ed., *John Donne: The Complete English Poems* (London and New York: Penguin Books, 1977); C. A. Patrides, ed., *The Complete English Poems of John Donne* (London: Dent, 1985); Ilona Bell, ed., *John Donne: Selected Poems* (London: Penguin, 2006).

outmoded and often inscrutable conventions of modernization; and they lacked consistent orthography. One of the more difficult decisions we made was to opt for modernized spellings, for the practical reason that we needed to be able to compare tokens across the database entries. (We could accommodate American and British spellings because most of their variations are formulaic.)

This problematically omits or marginalizes old-spelling sonnets. Editions like Emrys Jones's of Thomas Wyatt (1964) or Josephine A. Roberts's of Mary Wroth (1983) suffered from what I call "the *Amoretti* problem": the habit, particularly acute in Edmund Spenser's sequence, of antique and inconsistent orthography.[21] Regrettably, I never solved this problem but merely avoided it—by omitting Spenser altogether and transcribing old-spelling poets like Wroth and Wyatt as they were. The obvious solution would have been a stand-off markup scheme to regularize spellings, a step we reserved for later.

This raises another issue that I left unresolved, of cross-linguistic comparisons. The sonnet is international and multilingual. Zeroing in on English-language specimens tells us something about conventions that obtained in an English-speaking archipelago at Europe's edge, but nothing about the type—so long as any dataset omits Dante and Petrarch's Italian sonnets, of course, or vernacular adaptations like Luís de Camões's Portuguese sonnets. Thus, any dataset making claims to comprehensiveness has to include every sonnet in every language, and any toolkit making claims to universal applicability has to be language-agnostic. How else are we to determine how unique Camões's diction is, say, unless we can cross-reference his tokens with those that translate most directly into Shakespeare's? This is a tractable problem, but it is one that I have left for future scholars with better tools.

Why move from the trans-historical Norton anthology to the twenty-seven early sonneteers? This was the period when the English sonnet began, and these were its most prodigious authors. When Wyatt first translated Petrarch into English, he established a list of words and ideas that English sonnets contained. So just as Wyatt was the first to use words in a sonnet related to courtship, a later poet was the first to use words related to topics like marriage, death, or the sublime. Each subsequent sonneteer expanded that list.[22] This cumulative word list allows for one measure of how typical or

21. *The Poems of Lady Mary Wroth*, ed. Josephine A. Roberts (London: Louisiana State University, 1983).

22. The same could be said for any genre in prose or verse or drama, from the sermon to the city comedy. It is not about cumulative or mutual influences, but about the emergent common features delimiting

atypical a given sonnet is, in comparison to others. Consider the hypothetical scenario of two new sonnets by Philip Sidney in the 1580s, part of his *Astrophil and Stella* sequence. One contains only words that previous writers have used in their sonnets; the other contains only words that have never been used in sonnets before. (Include proper nouns in this thought-experiment, but ignore stopwords like articles, conjunctions, and prepositions.) Diction is just one imperfect measure of originality, but given these two sonnets we would be justified in calling the former typical and the latter atypical—or familiar and unfamiliar, if you prefer. We would rarely find such outliers, but they illustrate how diction could be one metric of typicality, and we would make this metric more subtle by including synonyms.

Consider Milton's sonnets, which the aforementioned critics characterize as atypical. What features of their language distinguish them from sonnets by other authors? This question of difference, like most others, is predicated on similarities. Formal and generic qualities of Milton's sonnets make them resemble others, particularly Petrarch's: fourteen lines of iambic pentameter with an Italian (octave-sestet) rhyme scheme; first-person complaints of love and descriptions of public occasions and private events. I can make this summary easily because Milton wrote just twenty-four sonnets: six in Italian, and eighteen in English. You can read them all in an hour, particularly with a good translation.[23]

So why would you need to distant-read them? Usually when we're talking about distant reading, the object is a vast collection of prose: every deposition given at the Old Bailey, for instance, or seven thousand British novels.[24] "Distant" reading is a deliberate antonym for "close" reading, the habit of paying sustained attention to localized language choices and effects. It gives you the capability to detect local text-features on a broader scale, which is why critics like Martin Mueller and Anupam Basu opt for the term "scalable reading."[25] "Scalable" suggests an extension, rather than a repudiation, of familiar worthwhile habits.

---

each genre and its variations.

23. Might I suggest John Milton, *The Complete Poems*, ed. John Leonard (New York: Penguin, 1998).

24. For the latter, see Franco Moretti, "Style, Inc.: Reflections on 7,000 Titles (British Novels, 1740–1850)," in *Distant Reading* (London: Verso, 2013), 179–210. In fact, the essay analyzes just the titles of these novels.

25. Martin Mueller, "Scalable Reading," 2020, sites.northwestern.edu/scalablereading/2020/04/26/scalable-reading/; Anupam Basu, "'Ill Shapen Sounds, and False Orthography': A Computational

The question is not whether you would want to distant-read Milton's sonnets in isolation, but what understanding they would yield in comparison to other sonnets.

The first question is which lemmas Milton uses in his twenty-four sonnets that appear in no other sonnets.[26] Many are proper nouns, owing to the contemporary occasions or classical allusions Milton favours as his subjects. They are, in descending order of frequency: cromwell, darwen, dunbar, babylonian, tetrachordon, gordon, macdonnel, cheke, cambridge, worcester, alcestis, cyriack, and piedmontese. (Like all other lemmas and tokens in the database, they are lower-case to permit comparisons.) Milton's sonnets tend to be about public or private occasions: deaths, dedications, critiques of his treatises ("tetrachordon"), and his blindness. They fall into a few categories: those on specific occasions, like the Roundhead's siege of London; and those praising specific people, like Henry Lawes, Margaret Ley, or the unknown lady of sonnet 9.

Far more illuminating, frankly, are the words that other sonnets use, that never appear in Milton's. Again, in descending order of frequency, they are these (now including their counts in the database):

> (sweet, 342), (loue, 282), (mee, 228), (beauty, 204), (hart, 198), (desire, 174), (fair, 167), (place, 167), (say, 165), (die, 145), (haue, 139), (faire, 130), (pleasure, 126), (nature, 114), (butt, 114), (selfe, 113), (flame, 104), (tear, 104), (pain, 103), (teare, 103), (wit, 100), (shee, 99), (cause, 97), (alas, 96), (forth, 93), (ill, 93), (tell, 90), (nott, 90), (burn, 89), (change, 89), (prove, 88), (dear, 87), (soule, 87), (self, 87), (glory, 86), (fall, 83), (breath, 83), (fayre, 79), (wind, 78), (sigh, 77), (write, 77), (wonder, 75), (griefe, 74), (neuer, 73), (yett, 73), (seek, 70), (passion, 69), (turn, 69), (sorrow, 68), (vnto, 68).

This list betrays a few clear limits of our process. Wyatt and Wroth's old-spelling words should be disregarded, because the process didn't disambiguate between

---

Approach to Early English Orthographic Variation," in *Early Modern Studies after the Digital Turn*, ed. Laura Estill, Diane J. Jakacki, and Michael Ullyot (Toronto: Iter Press, 2016), 167–200.

26. A "lemma" is the equivalent to a word-token's dictionary headword, so "ran" and "running" share the lemma "to run"; lemmas permit more refined comparisons between texts, and higher-order processes like lists of their parts of speech.

two spellings of the same word ("love" and "loue" or "self" and "selfe"). But it also reveals that Milton never uses words that seem associated with the syrupy love-sickness that we sometimes see in Elizabethan sonnets (sweet, beauty, desire, fair, pleasure), nor with Petrarchan suffering (die, flame, pain, alas, ill, burn).

And yet these lemmas are only from the eighteen sonnets that Milton wrote in English; as I said, he also wrote six Italian sonnets. They are his first sequence, written during his Italian travels of the late 1620s. Milton has clearly read his Petrarch, and is keen to imitate him; in sonnet 3, he writes (in John Leonard's translation) that "Love wakens on my quick tongue the strange flower of a foreign language."[27] The Italian sonnets are Milton's return to a source of vernacular sonnets, to learn the form and its conventions *ad fontes* rather than from his English contemporaries. The poet who scorned the barbarous custom of rhyme to justify using blank verse for *Paradise Lost* opted for an older, foreign model for this lyric form.

Accordingly, Milton's English-language sonnets differ markedly from those of other vernacular authors. Take just the example of religious sonneteers. Unlike Donne or George Herbert, Milton uses other forms to address God or to describe prayer, baptism, sin, redemption, or other devotional topics. Here are the lemmas, again with frequencies, that Donne's thirty-seven sonnets use that never appear in Milton's eighteen:

> (oh, 22), (die, 21), (sleep, 18), (last, 17), (burn, 13), (mourn, 12), (flesh, 11), (take, 11), (hell, 10), (black, 9), (weak, 9), (pain, 9), (despair, 8), (drown, 8), (blow, 8), (body, 8), (cannot, 8), (tear, 8), (begin, 8).

As you might expect from his preoccupation with suffering, Donne uses many more lemmas related to mental weakness, fleshy frailty, and divine punishment than Milton does—or at least, than Milton does in his sonnets. Similarly, the self-mortification of Herbert's seventeen sonnets distinguishes their lemmas from Milton's with three (burn, cannot, and die) in common with Donne:

> (burn, 6), (since, 6), (cannot, 6), (poor, 5), (turn, 5), (take, 5), (beauty, 5), (seek, 5), (sound, 5), (evry, 5), (wit, 5), (unto, 4), (die, 4), (ink, 4), (invention, 4), (low, 4), (suit, 4), (birth, 4), (flame, 4), (church, 4).

27. Milton, *Complete Poems*, 32.

We also see Herbert's structural arrangement of his sonnets around the architecture of the church, and above all his stress on the poet's own uses of wit, invention, and ink. Borrowing a term from artistic depictions, these lists are like negative space—revealing Donne's and Herbert's sonnet diction in contrast to Milton's. We must not overinterpret them, something that I risk here. The only thing these lists tell us is that in Milton's eighteen vernacular sonnets, he didn't opt for words that other poets used in theirs. That only means that he made narrower use of the sonnet, and that Donne and Herbert addressed topics and implied readers that Milton reserved for other poetic types.

We need no ghost in the machine to tell us this. The authoritative critics Anna K. Nardo, Richard Streier, and Barbara Lewalski established the orthodoxy long ago that Milton's sonnets are atypical.[28] I am affirming that orthodoxy with quantifiable evidence. To object to the evidence of digital text-analysis "but we knew that already" misses the point. Many orthodoxies are valid. Determining which ones are worth preserving is the point.[29]

Modelling the sonnet

The sonnet is a poetic type characterized by intersecting formal and generic conventions that shifted and accumulated through time. Its model, therefore, is a cumulative and multidimensional set of ranked conventions of varying weights: so much for meter, so much for diction, and so on. This departs from the Neoplatonic sense that Milton uses for models in *Paradise Lost*, when God mocks earthly cosmologists' efforts to "model heaven" (8.79). Colin Burrow describes this as "an attempt at earthly imitation of a heavenly truth," akin to

28. See Anna K. Nardo, *Milton's Sonnets: The Ideal Community* (Lincoln: University of Nebraska Press, 1979); Anna K. Nardo, "Milton and the Academic Sonnet," in *Milton in Italy: Contexts, Images, Contradictions*, ed. Mario A. Di Cesare (Binghamton, NY: Center for Medieval and Renaissance Studies, 1991), 489–503; Richard Strier, *The Unrepentant Renaissance: From Petrarch to Shakespeare to Milton* (Chicago: University of Chicago Press, 2011), dx.doi.org/10.7208/chicago/9780226777535.001.0001; and Barbara K. Lewalski, "Contemporary History as Literary Subject: Milton's Sonnets," *Milton Quarterly* 47.4 (2013): 220–29, dx.doi.org/10.1111/milt.12055.

29. As Matthew Jockers has observed, there is a curious, unexamined assumption among literary critics that complication and contestation are more worthwhile than confirmation (*Macroanalysis: Digital Methods and Literary History* [Champaign: University of Illinois Press, 2013], 31), dx.doi.org/10.5406/illinois/9780252037528.001.0001; one need only consider the replication problem in psychology to see that this assumption is particular to our discipline.

an architectural plan directing builders who realize it. For Milton, a model was a set of "patterns, or replicable formulae, for composing works."[30] The model I am building works in the opposite direction, taking measurements of sonnets as they exist in the world; it is closer to a network of jostling criteria than to a Platonic form.[31]

Yet a model of the sonnet is the apotheosis of every poet's idea of this poetic type, every notion of its formal and generic characteristics. They will overlap, but they will also be particular to different poets' learning and reading experiences. No poet can read every sonnet ever written, to form a mental model of it—just as no critic can, to compare a given sonnet with their mental model of sonnets. Our models are never objective, but they are inter-subjective, distilled from the texts we've encountered. (For "model" here you could substitute "grasp" or "heuristic.") We can also apply models at different scales. I might have a mental model of one Milton sonnet, or of all of his sonnets, or of the various genres they exhibit. But the higher up this scale, the less true and less useful my model becomes—useful, that is, for novel instances that I encounter, and useful for you and me to discourse about our shared knowledge.

Willard McCarty describes models as instrumental, in the sense of facilitating inquiry and knowledge; they are "temporary states in a process of coming to know" a phenomenon or a natural occurrence, enabling what Piper calls "surrogative reasoning."[32] Drawn from the history of the sciences, models are limited in their *intrinsic* power to represent, but enabling in their *extrinsic* power to drive interpretations: think of the model of a chemical structure or of the solar system, a representation lacking many details but providing the essential structures that we need to intuit both natural phenomena. Models are always reductive, but the rationale for using them is to replace our confident, false formulations with more humility, more susceptibility to testing, to error. For this reason, Richard Jean So suggests that models reveal the illuminating

30. Colin Burrow, *Imitating Authors: Plato to Futurity* (Oxford: Oxford University Press, 2019), 314–15, dx.doi.org/10.1093/oso/9780198838081.001.0001.

31. This dichotomy is Moretti's, in "Slaughterhouse."

32. Willard McCarty, *Humanities Computing* (Houndmills: Palgrave Macmillan, 2014); Piper, "Think Small."

departures from norms, in this case the sonnets that lie well outside the typical ones.[33]

Consider the edge-case sonnets of Donne. Although some of Donne's sonnets take a conventional Shakespearean form, there are provocative outliers. Features of those sonnets suggest his view of this poetic type as porous and permeable, or more so than Milton's Italian apprenticeship encouraged. Donne's own apprenticeship was in verse letters, a flexible occasional genre of direct address that sometimes exhibits formal and topical features that make it quite sonnet-like. For the purpose of our database, my students and I included every fourteen-line verse letter of Donne's, a decision that arbitrarily if provisionally defined the sonnet as a form; any disputable sonnets that were thus included would, theoretically, be vastly outnumbered by indisputable sonnets. This distinction returns us to the critical orthodoxies that began this argument. They underscore the seeming paradox of a method that uses orthodoxies to obviate themselves.

This method is not reframing a problem, like going west to go east. Rather, it is a concession to the weight of expert opinion. Categories like the sonnet can accommodate exceptions only after similarities establish the rule, the statistical norms of formal features. We begin with the assumption that Donne's twenty-six *Holy Sonnets*, using a Shakespearean rhyme scheme, are the norm, and his eighteen-line "Sonnet. A Token" is an outlier. Formally, it is a Shakespearean sonnet with an extra quatrain. Topically, it fits the generic conventions of a sonnet: a speaker's first-person address to his beloved, cataloguing the inadequate tokens she might send of her love, culminating in his turn (or *volta*) away from them in a conclusive couplet. Except for the extra quatrain, it is a conventional sonnet. Although there's no reason to think that Donne designated it a sonnet, the compiling editor of his 1635 collection *Songs and Sonnets* gave it this title. In this collection, at least, Donne's sonnets have a topical or generic definition rather than a formal one.

Returning to his verse letters, they present more difficult quandaries for the database-building human reader. Some, like "Witchcraft by a Picture," meet the line-number convention but violate both the rhyme-scheme and topical

33. Richard Jean So, "All Models are Wrong," *Publications of the Modern Language Association of America* 132.3 (2017): 668–73, dx.doi.org/10.1632/pmla.2017.132.3.668.

conventions; they are not sonnets.[34] But others follow a hybrid Petrarchan-Shakespearean rhyme scheme, ABBA ABBA CDDC EE ("Thy friend, whom thy deserts to thee enchain") or a hybrid Petrarchan-epistle scheme ABBA CDDC EE FF GG ("Kindly I envy thy song's perfection").[35] On the basis of their "vocabulary" and "conceit[s]," respectively, the 1967 Clarendon Press editor Wesley Milgate designates both as sonnets.[36] They fit Milgate's mental model of a Petrarchan sonnet.

The alternative to such authoritative declarations and critical intuitions is an objective model, built on the same cumulative procedure by which Milgate and others develop mental models. Two important differences are scale and randomization: the process addresses far more sonnets than a critic could read, or could retain in their memory; and it is indifferent to authorial attributions or other human valuations. It would unfold something like this: Take two sonnets as examples, at random; compare their quantifiable features. Then add a third sonnet; fold its quantities into the model. And so on for a thousand sonnets. And then for ten thousand. What you end up with is a statistical score for each text's typicality. But to do this, you need to start with a verified set of known sonnets, tagged as such. They need to be regularized in some way: all in the same language (English), all using the same spellings for tokens (modern), all machine-readable text with spaces between words, and with hard returns between lines. Then you can lemmatize and do other processing to their tokens and characters.

A universal model of sonnets, starting with those in the English language, has manifold advantages over what human critics already do naturally. It can tell me if my intuitive grasp of Milton's difference from Herbert or from

34. So claims Ilona Bell, Donne's editor, in *Selected Poems* (London: Penguin, 2006), 123–25. Other fourteen-line verse letters that are not sonnets are the three addressed to T. W. (probably Thomas Woodward, b. 1576): one ("Haste thee harsh verse as fast as thy lame measure") consisting entirely of couplets; the others ("Pregnant again with th'old twins, Hope and Fear," and "At once, from hence, my lines and I depart") of four tercets and a closing couplet.

35. The first Donne addresses to C. B. (Christopher Brooke, ca. 1570–1628); the second to R. W. (Rowland Woodward, 1573–1636/7).

36. "The vocabulary of the [former] poem comes from the conventional Petrarchan stock. […] Donne here takes the trouble to achieve a genuine sonnet" (John Donne, *The Satires, Epigrams and Verse Letters of John Donne*, ed. W. Milgate [Oxford: Clarendon Press, 1967], 215, dx.doi.org/10.1093/actrade/9780198118428.book.1). The latter "is a sonnet, built on the conceit of the four elements" (219).

Donne is supported by diction—even if lemma choices are merely one form of evidence. It can not only support or discredit my intuitions but also make them more nuanced by localizing them in individual lemmas. It can confirm or deny Milgate's assertions that these two verse-letters are sonnets, but not those others. The benefits of a quantifiable model of the English sonnet will be manifold: a vastly expanded canon of sonnets, beyond the most anthologized specimens; a categorical definition of sonnets that exactly specifies the degree to which it is formal or generic; a sense of its historical development through time and across periods, authors, and anglophone poetic cultures; and a knowledge of sonnet subgenres and just how they depart from each other. In sum: an empirical grasp of one poetic type that accounts for its multivariate complexity—not a flattening of differences, but a set of metrics to measure differences.