Introduction: Computation and Digital Text Analysis at *Melville's Marginalia Online*

Christopher Ohge, Steven Olsen-Smith

Leviathan, Volume 20, Number 2, June 2018, pp. 1-16 (Article)

Published by Johns Hopkins University Press
DOI: *https://doi.org/10.1353/lvn.2018.0017*

Introduction:
Computation and Digital Text Analysis at Melville's Marginalia Online

CHRISTOPHER OHGE and STEVEN OLSEN-SMITH
University of London and Boise State University

In an erased marginal comment on Shakespeare's character Parolles, the dissembling rogue in *All's Well That Ends Well*, Herman Melville invoked the most unexceptionable of mathematical formulas to reflect darkly on the abiding nature of human depravity:

> As 2 & 2 made 4 in Noah's time, as now,
> so man [?figures] ever. Here we have a
> character very common in the Rail Road
> Car of the [?most mighty] nineteenth century.[1] (***Shakespeare's Dramatic Works*** 2.406)

Roughly forecasting the subject of Melville's tenth book, *The Confidence-Man*, the recovered annotation also reveals a seemingly incidental but nonetheless significant penchant on Melville's part for quantification. Quantities appear frequently in *Moby-Dick* and in other writings by Melville, whose fascination with numbers was demonstrated recently in the pages of this journal by Zachary Turpin. Numbers are not scarce in the author's marginalia either. "There are 75 folio volumes in that," Melville observed of Jaques's words in *As You Like It* (2: 284.26–27). In light of this predisposition to unite the quantifiable with the profound, the *Melville's Marginalia Online* (*MMO*) staff offer in the following essays digital text analyses of Melville's marginalia in his surviving copies of Homer, Shakespeare, and Milton. Methodologically unprecedented, and conducted in an integrative spirit, the essays combine distant with close reading, exploring hitherto unknown or under-appreciated evidence of Melville's engagement with these three major predecessors.

Quantitative and qualitative digital scholarship have led to a mixture of promising results and disillusionment, with increasingly vocal proponents and detractors. Opposing conceptions of the field are apparent in Timothy Brennan's recent article, "The Digital Humanities Bust," and the

Vol. 20.2 (2018): 1–16

vigorous defense it provoked from many digital scholars (for examples, see responses by Mandell and by Bond et al.). While the fairness and accuracy of Brennan's characterizations of digital practitioners' various activities is open to question, he makes a valid point that quantitative studies could more explicitly demonstrate the significance of their findings. High-profile studies to date have mainly addressed broad swaths of social literary output defined by regions and decades––for instance, the thousands of British novels published over a given century. Digitized archives now make such approaches possible and offer an exciting and potentially transformative development in literary studies. On the other hand, the stubborn disconnect between quantity and literary meaning can hamper such approaches, presenting conceptual shortcomings even for macro-analyses performed on data of a comparatively much smaller scale. As Brennan puts it bluntly, "The significance of the appearance of the word "whale" (say, 1,700 times) is precisely this: the appearance of the word 'whale' 1,700 times." But whereas no act of counting can substitute for critically interpreting a literary work, the usefulness and significance of word frequencies––both on their own and in relation to each other––are heightened in the study of marginalia. Indeed, the approaches demonstrated in the following essays indicate that an author's record of marginalia and lifetime of reading constitute a valuable big data set in its own right, albeit a fairly small one in the context of research computing. When the author is Melville, moreover, with a substantial record of marking his books and incorporating a range of sources in his own writing, the data can reward analysis in hitherto unprecedented ways. Only a machine can quickly compute word counts of selected content Melville marked in the plays of Shakespeare's *Dramatic Works,* or in each book of *The Odyssey,* or in *Paradise Lost*. Only a machine can stack a series of graphs showing lexical variety in Melville's markings or instantaneously classify by positive and negative sentiments enormous amounts of words that appear most frequently in passages he marked in multiple texts. Melville himself was devoted to the art of rhetoric and explicitly lauded both the "short, quick probings" and encyclopedic forays of great writers ("Hawthorne and His Mosses" 244). By using machine-derived word counts and visualizations to explore the thematic aspects of Melville's reading, the following essays demonstrate new means for assessing his influences, source use, and conceptual development. Coupled with close reading, so-called "distant reading" techniques can yield surprising revelations that alter or enhance our convictions about Melville and his sources.[2]

While the following three essays focus separately on Melville's marginalia to Shakespeare, Milton, and Homer and use different methods, quantifying and

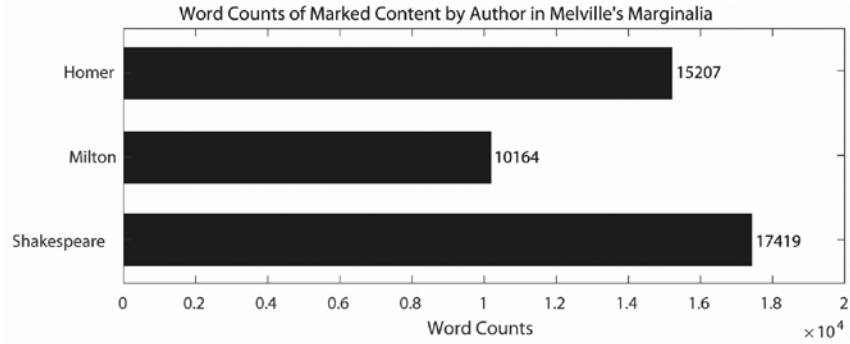

Fig. 1. Bar graph comparing total word counts of marked passages in Homer, Milton, and Shakespeare.

visualizing the three data sets in relation to one another can help to introduce some of the methods employed for each. These essays forecast the expansive approaches to Melville's reading that will come about as digitization continues at *MMO*. For example, word-count graphs for marked content in Melville's copies of Shakespeare, Milton, and Homer display, at a glance, the extent of his engagement with their works. In addition to offering rapid comprehension and insights that facilitate user access to and engagement with evidence, these visualizations generate new questions that may not occur to a reader who simply browses through each instance of marginalia. Chiefly useful for the evidence they offer of Melville's marginalia in their totality, graphs and tables that compare word or character counts offer a sweeping view of the content Melville marked and annotated in his library. As demonstrated in Fig. 1, Melville marked more content in Shakespeare than in Homer and Milton. Whereas that fact might be said to follow predictably from the respective number of volumes in each set he owned (seven volumes in *The Dramatic Works of William Shakespeare*, four in *The Odyssey* and *The Iliad* of Homer, and two in *The Poetical Works of John Milton*), it may be surprising to see that the Homer word count comes close to the total in Shakespeare. In each of the following essays, a graph of the word counts per section (per play in Shakespeare, per work and by book in Milton and in Homer) enables analyses of where and to what extent Melville devoted marginalia.

Lexical richness calculations in the following essays derive from hapax analysis. Hapax analysis renders a percentage based on the number of words that occur only once (called *hapax legomena*) divided by the total number of words. It is arguable that hapax legomena or unique word clusters may not be as significant to uniqueness as the repetition of common (or groups of related)

words; for example, John F. Burrows argues that the styles of authors come from common words (i. e., articles and prepositions) that "constitute the underlying fabric of a text, a barely visible web that gives shape to whatever is being said" ("Textual Analysis").[3] However, in the present data set a variety of calculations (from simple character counts to hapax percentages) can help to identify repetition, brevity, and unique word clusters. The fragmented, specialized nature of marginalia—emphasizing ideas—requires more attention to unique words and collocates than the average text analysis project.

The ratios in the lexical uniqueness graphs illustrate the differences among individual markings; that is, lexical richness is recognized by comparing and contrasting the heights of the bars in a given graph. Each bar represents its own marking instance and the independent value of its content compared to the whole. Where and under what conditions did Melville attend to word variety or consistency in his markings? to repetition? or to substantive ideas with unique vocabulary? These visuals illustrate a pattern of unique word occurrences in brief passages or repetition in lengthier ones. Whereas lexical richness calculations involve a measure of risk and imprecision (see Baayen 258), hapax analysis identifies a large percentage of the words marked by Melville due to his tendency to underline words and expressions he found meaningful; and these underlined passages, often of unique words, frequently result in maximized percentage values in the graphs. Identifying the passages of higher and lower hapax percentages from a distance aids the close study of marked content. For example, the first marking in Shakespeare's *The Tempest* returns a lexical variety value of .88, meaning that each word type is used an average of .88 times (see the first bar in Fig. 2, which shows a calculation with hapax percentage). In other words, the passages contain mostly unique words: "That this lives in thy mind? What seest thou else / In the dark backward and abysm of time?" (only one word here—*in*—occurs more than once).

The lower percentage values in the graphs draw attention to sections with less variety of word usage, suggesting ways quickly to find brevity or repetition (see Jockers 61–64). Exemplifying the latter, a passage marked by Melville in Book 8 of *Paradise Lost* has a low level of lexical variety (the second-to-last bar, as seen in Fig. 3):

> To whom the angel with a smile that glow'd
> Celestial rosy red, *love*'s proper hue,
> Answer'd. Let it suffice thee that *thou* know'st
> Us happy, and without *love* no happiness.
> Whatever *pure thou* in the body enjoy'st,
> (And *pure thou* wert created,) We enjoy
> In eminence, and obstacle find none

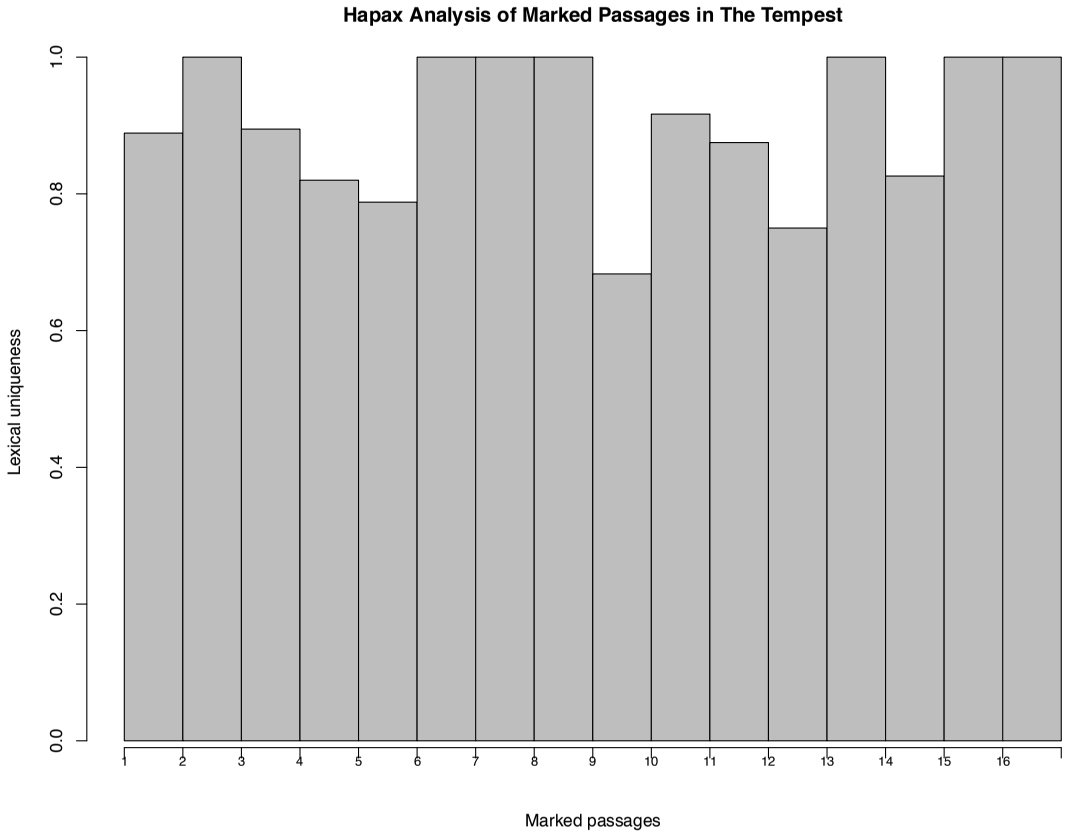

Fig. 2. Graph of unique word percentages (lexical uniqueness) in Melville's markings of *The Tempest*.

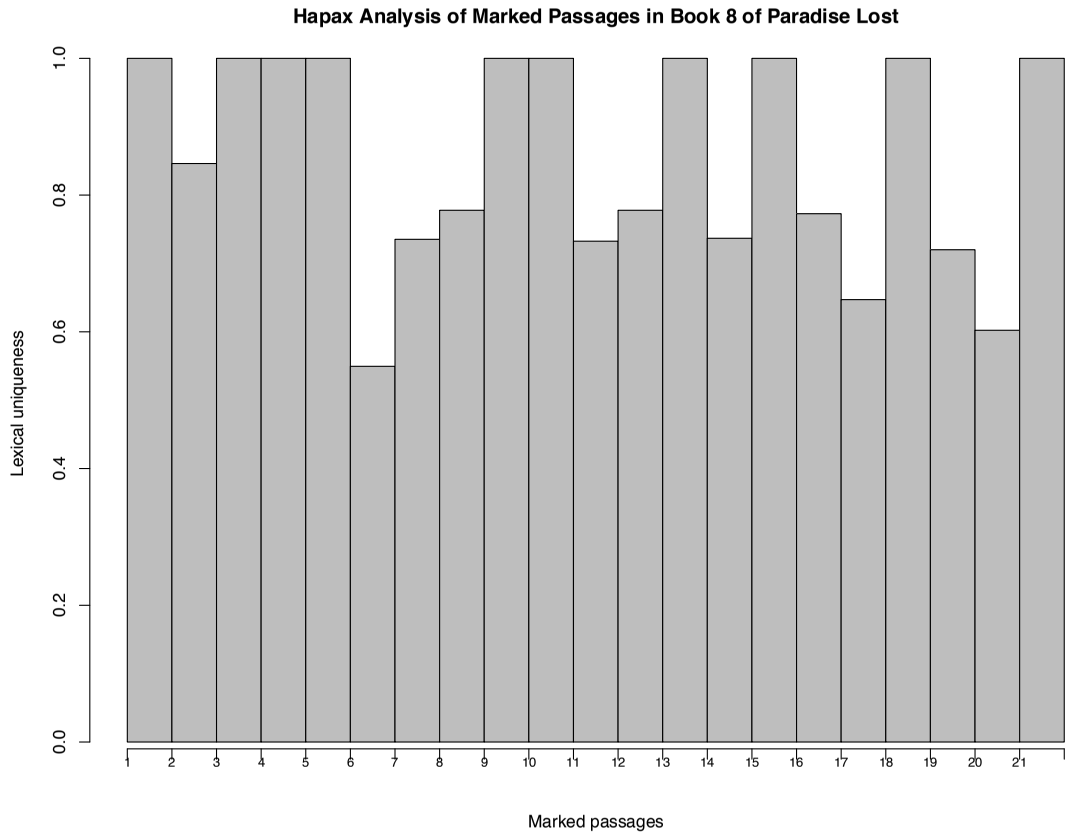

Fig. 3. Graph of unique word percentages (lexical uniqueness) in Melville's markings of Book 8 of *Paradise Lost*.

> Of membrane, joint, or limb, exclusive bars:
> Easier than *air* with *air*, if spirits embrace,
> Total they *mix*, union of *pure* with *pure*
> Desiring; nor restrain'd conveyance need
> As *flesh* to *mix* with *flesh*, or *soul* with *soul*."
> (1: 275; emphases added to show repeating substantives)

The repetition is both literal (in the case of "pure," "love," "happy," and "soul") and conceptual, showing that "flesh" and "body" are accompanied by the related words "membrane," "joint," and "limb." The form of address also changes from the repeating "thou" to a powerful, singular "We" after the parenthetical pause. Attending to simple repetition leads to the discovery of a more graceful form of repetition and of thematic significance through the use of synonyms, demonstrating the value of consulting these simple linguistic analyses to examine multiple stylistic phenomena. Approaching a marked passage through its lexical variety obliges readers to consider the passage linguistically, to notice its repetitions, collocates, word lengths, and parallel constructions. Such linguistic consideration can bolster attentive reading; indeed, the reading experience is fundamentally changed—and for the better—by acknowledging the linguistic-numerical facts underlying the passages Melville chose to mark while he read.

The variety of vocabulary (termed "lexical uniqueness" in the following essays) in Melville's marginalia suggests how the linguistic character of marked passages might compare to overall lexical varieties in the works he read in their totalities. Mainly useful for comparing the lexical character of full texts among different authors (a different approach from the comparative analysis of marked content), the lexical analysis in Fig. 4 comes with a caveat: variety is generally lower as the total word count rises, which is why the seven-volume Shakespeare set has the lowest value in the graph. The higher uniqueness levels in the content Melville marked reveals that he was generally attending to breadth of expressions and ideas rather than repetitive words or phrases. Nevertheless, his markings constitute a smaller set of specialized words and expressions, so they should show more lexical variety. Even though Melville's marginalia to Shakespeare have the least variety of the three marked sets taken up in the following essays, the graph shows that the lexical values in the Shakespeare markings differ significantly from the variety in the total text, whereas the values for Homer are comparatively closer, suggesting that Melville marked more unique words in Shakespeare relative to the whole text than he did in Homer.

In addition to lexical uniqueness, word sentiments from the marking evidence can be ordered and visualized to initiate analysis of key words in context

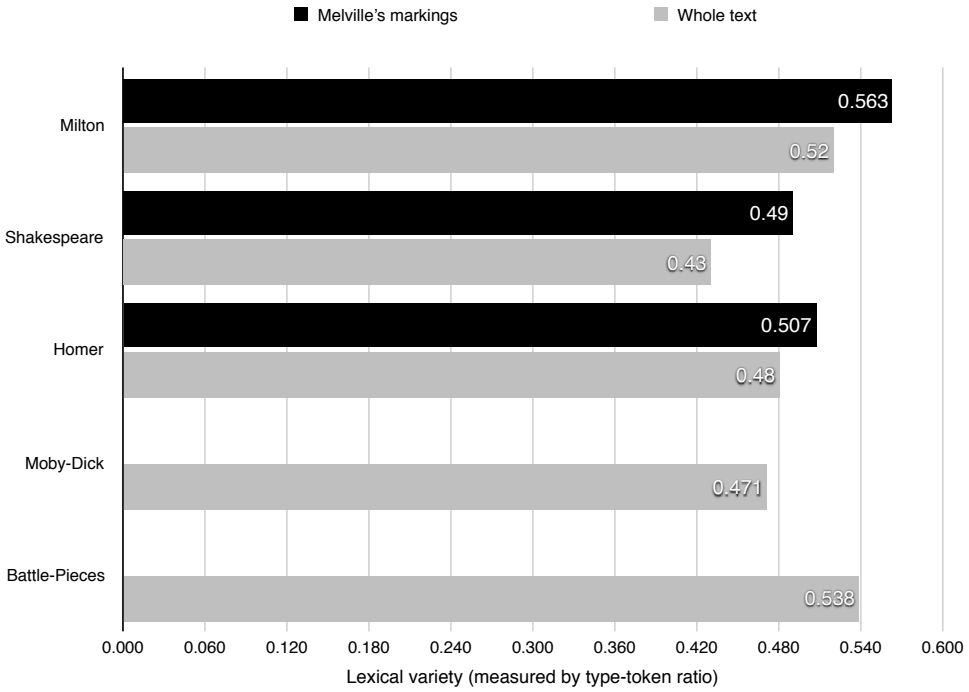

Fig. 4. Bar graph of type-token ratios in Melville's markings compared to the whole texts of Melville's reading of Homer, Milton, and Shakespeare. The bottom of graph shows total type-token ratios of Melville's *Moby-Dick* and *Battle-Pieces*.

(see Fig. 5). Drawing on Silge and Robinson's "tidytext" package in R, sentiment analysis shows the frequency of positive and negative words in a given data set. While there is a greater net number of negative words in Melville's marked content in the writings of all three authors, the sentiment graph shows that he noted a small group of positive words with more frequency than the negative ones. The negative words are more variable as well as more numerous. The bar graphs of the twenty most frequent positive and negative words allow one to posit new questions about frequently-used words and their implications within marked passages. What can be inferred from the high frequencies of negative terms and the concentrated appearance of select positive terms? Whereas Fig. 5 conglomerates terms in Melville's marginalia to Shakespeare, Milton, and Homer by frequency and sentiment, the lexical dispersion plot in Fig. 6 shows where the most frequent positive and negative words occur across Melville's marginalia in all three authors. Whereas "fear," "fall," "great," and "love" appear with some consistency throughout Melville's marginalia to all

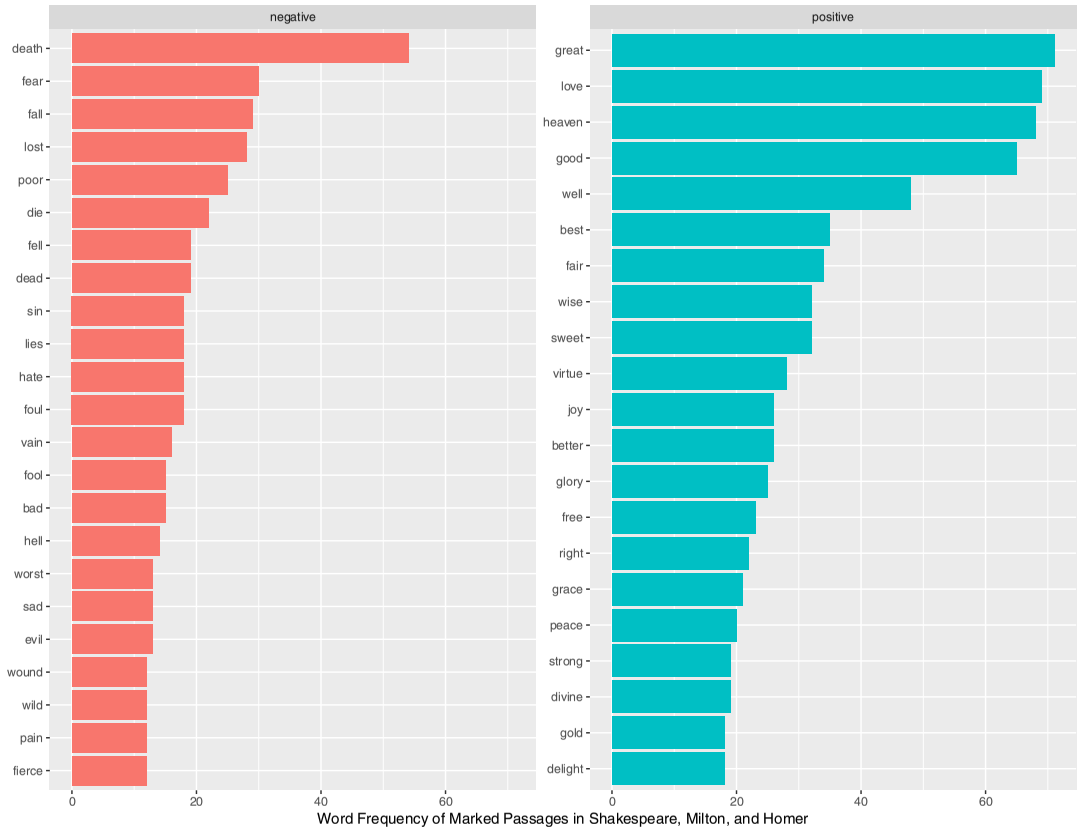

Fig. 5. A graph showing the most frequent sentiment words in Melville's marked passages in Homer, Milton, and Shakespeare.

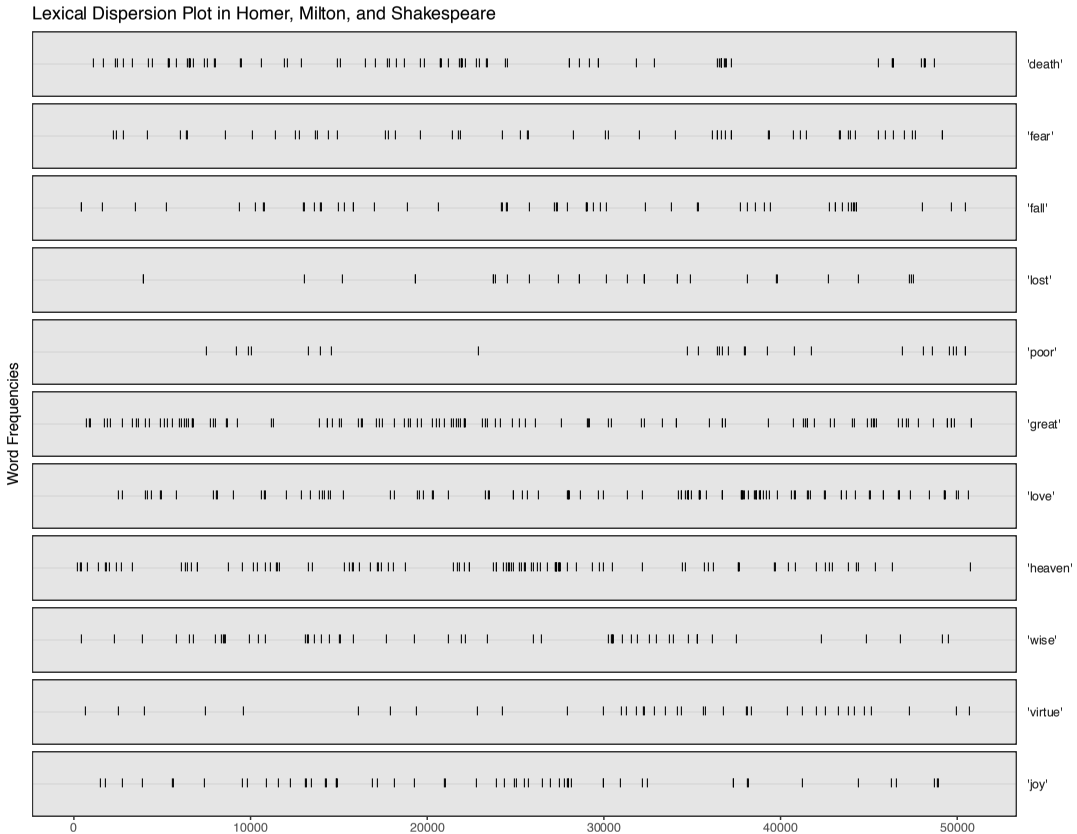

Fig. 6. A lexical dispersion plot showing the occurrences of high-frequency terms in the Homer, Milton, and Shakespeare markings.

three authors, "virtue" features more in Milton and Shakespeare and "wise" appears more concentrated in Homer and Shakespeare. Also noteworthy is the heavier frequency of apparently positive-aspirational terms ("joy," "heaven") in Homer and Milton. This can provoke further word searches for co-occurring sentiment words in areas of marked text. It is important to stress that the most frequent sentiment words are a fairly small subset of the total corpus, and focusing the data on sentiment words can draw attention to some concepts that might be overlooked by simply engaging with high frequency terms.

The overall picture of sentiment data, while useful, is merely a starting point for closer analysis using the keyword search tool at *MMO*. The next step would be to provide one more level of quantifiable specificity: how many of the total instances of "love" or "heaven" or "great" are complicated or undermined by the context? As is shown in detail in the Shakespeare essay, the high-frequency positive word "love" often appears in a negative context. Another suggestive word high on the positive list, "great" is well-distributed throughout the three authors: 36 times in Homer, 32 in Shakespeare, and 12 in Milton. Yet a word search of "great" in the Shakespeare markings alone on *MMO* shows the different senses of the word, ranging from simple intensifiers ("a great deal," "too great an act") and emphases of negative ideas ("great guilt," "great affliction"), to strongly positive modifiers ("great Bolingbroke," "great reason," "great self"). Further analysis of sentiment words in different authors will reveal where the sentiment distinction is useful and in some cases (as with "love") will indicate where a sentiment word is the locus beneath which a more complex meaning may lie.

The ability to visualize frequencies and clusters in the marginalia and to examine key terms in context through the search tool at *MMO* provides new access to the evidence of Melville's reading. Plans for future technical upgrades at the project include dynamic visualizations that will link directly to the holograph evidence from which they derive. Additionally, the stylometry library in R called "stylo" promises to complement marginalia analysis with distant readings of the whole texts that Melville read. Built into stylo's complex algorithms are some data calculations that already have been marshalled for the analysis of marked text: average word length, lexical uniqueness, the most common function words, and 2 or 3 n-grams (word pairings or triples). Fig. 7 shows a dendogram generated by stylo that compares the full texts of Shakespeare, Milton, and Homer to Melville's own literary output. Stylo's accuracy shows in the visualization, which groups most of Melville's works near each other in what might be thought of as a stylistic family tree. The greater proximity of Melville's writings to Shakespeare's shows that, from a linguistic perspective, Melville's style is a closer cousin to all of Shakespeare's plays than to Homer or Milton.

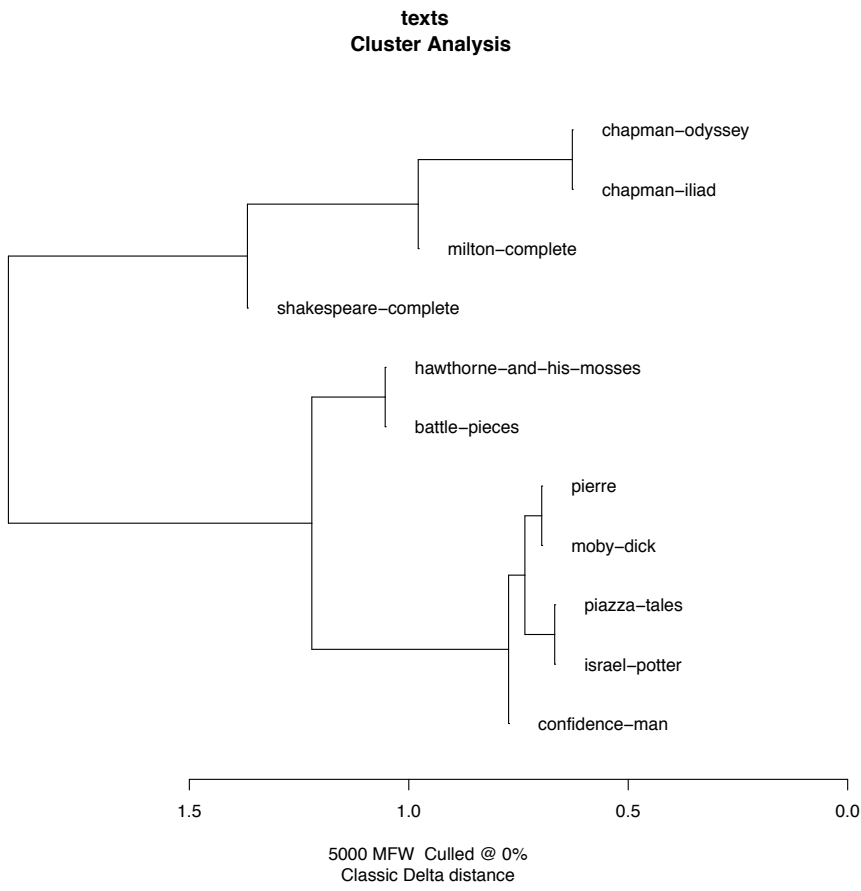

Fig. 7. A dendogram (tree), created with stylo, showing the similarities between Melville's early works and the whole texts of Homer, Milton, and Shakespeare.

Chapman's rendering of Homer is farthest from Melville, perhaps owing to the fact that Chapman was working from a Latin translation. These data points might get us beyond the grey area of genius: we know that Shakespeare, Milton, and Homer influenced Melville, but how and why do stylistic alignments and deviations occur?

Stylo can identify the most frequent function words in the whole texts of Homer, Milton, and Shakespeare and set this data against Melville's works. Burrows argues in *Computation into Criticism* that the most common words are part of an author's signature style. For example, it is important that Melville, like Shakespeare and unlike Milton and Homer, favored the first person over the third person. What stylo can do is to gauge the similarities (down

to the last minutiae that human readers are unlikely to notice) in Melville's writings to the works he was reading, suggesting new investigations of stylistic comparisons. This analysis of simple function words can be complemented with unique word differences. As is seen in Fig. 8, stylo can also group the most distinctive words in Melville's reading as compared to his own words in his works. Among the most distinctive words (other than function words) in the whole texts of Homer, Milton, and Shakespeare, "honour," "grace," "son," and "father" stand out, suggesting themes of virtue and legacy. On the other hand, some words in Melville's writings that diverge the most from these readings—such as "seemed," "moment," "like," and "something"—seem related to perception. These discoveries are catalysts for new analytical directions. Going forward, *MMO* plans to use stylo for rapid calculations and visualizations of linguistic phenomena for whole texts of Melville's reading in connection with scholarly electronic editions of his writings at the *Melville Electronic Library*.

Marginalia cited in the following studies have been transcribed and encoded in extensible markup language (XML) and delineated according to procedures explained at the "Policies" page of *MMO*. The quantified data include only printed text marked by Melville, not the content of annotations written by him in the margins of the three sets or autographs and notations inscribed by him on end leaves. Content designated as marked corresponds to *MMO*'s printed line designations in the left sidebar textual apparatuses of the digital copies of *The Dramatic Works of William Shakespeare* (Sealts no. 460), *The Poetical Works of John Milton* (Sealts no. 358b) and George Chapman's translations of *The Iliad* and *The Odyssey* (Sealts nos. 277 and 278). As explained in "Policies," apart from straightforward instances of underlined content, the exact extent of marked content is not always clear in Melville's marginalia, since marginal checkmarks and even scorings often raise some ambiguity about the beginning- and end-points of the passages that concerned him. *MMO*'s method in such cases is to err on the side of inclusion, trusting that a portion of superfluous words in the data set is preferable to the omission of words intentionally marked by him. The numbers of words construed as "marked" in the following analyses correspond exactly to their quantities in the XML markup, which can be downloaded at the project's github page and verified in individual instances by using the search tool at *MMO*.[4] When tallying distinct instances of marking, however, our method is to differentiate between embedded marginalia and their fuller marked contexts. An underlined passage, for instance, should be considered part of, but also apart from, the larger scored passage in which it is situated. Corresponding to the parent-child hierarchy of XML nodes, embedded and contextual instances of marginalia are treated as adjacent but distinct instances of marking.

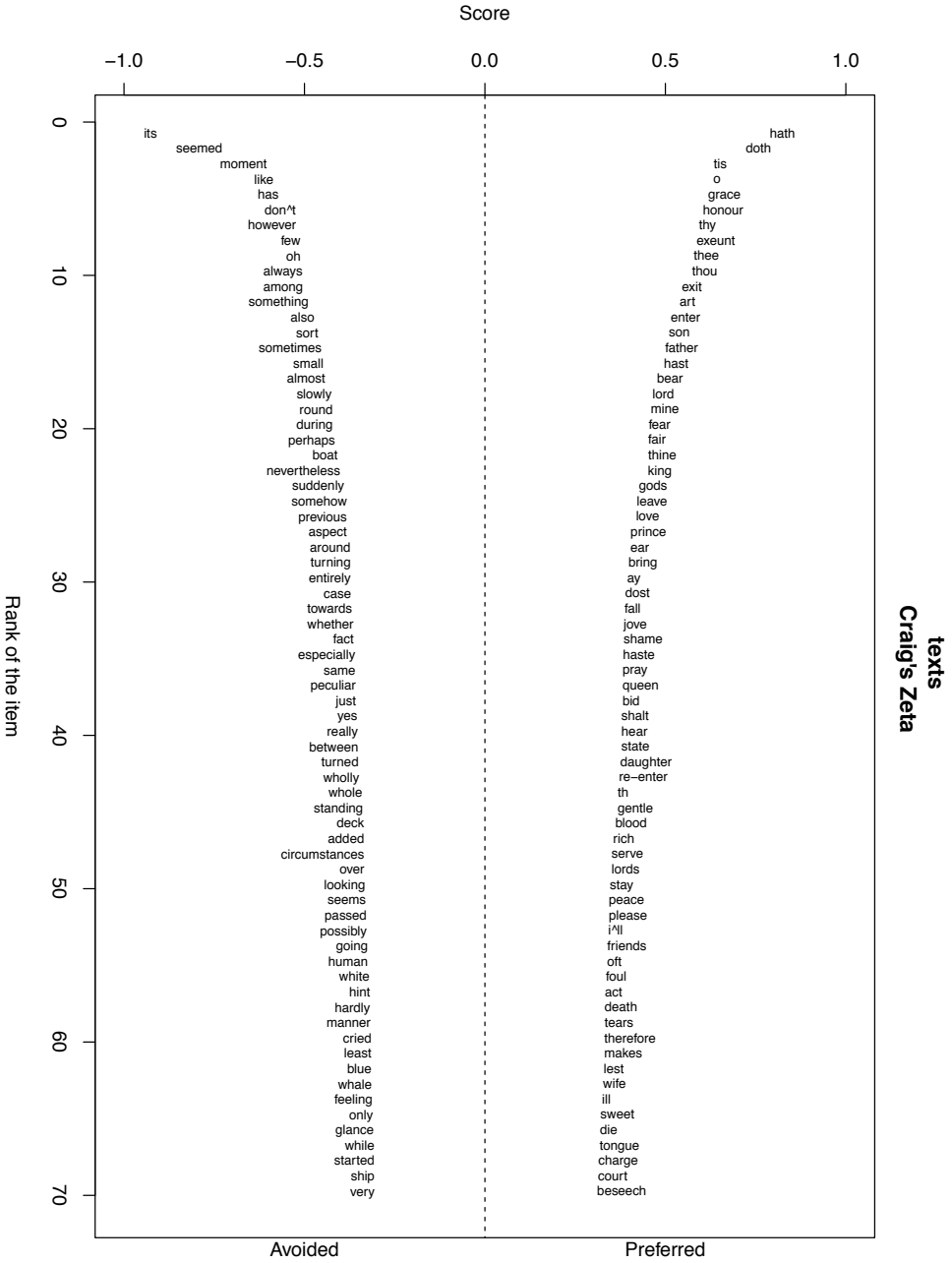

Fig. 8. A word plot, created with Craig's Zeta, showing the word preferences between Melville and Homer, Milton, and Shakespeare.

The following essays show how the digital text analysis of Melville's marginalia reveals new aspects of his reading that would have been impractical—if not impossible—using traditional methods. The aim is to make source study more holistic, examining as many angles as possible to understand Melville's reading and writing practices. Even if some of these distant reading techniques verify arguments already based on more traditional methods, the linguistic facts can now be cited to support those assertions. Of course, factual-quantitative findings need to be questioned as vigorously as critical opinions based on careful readings. Instead of getting trapped in what Adam Hammond calls a "double-bind of validation," these digitally-enhanced opinions lead back to the original documents to make the case for their significance, navigating between distance and depth without committing uncritically to either in isolation.[5] Our nuanced structures of reading always benefit from more flexibility, expansion, and curiosity, all of which can be achieved with the help of digital methods presented in the following pages.

Notes

The guest-editors thank Samuel Otter and Brian Yothers for precise feedback on the analyses and for their guidance with the array of figures in the essay cluster. Dennis C. Marnon, and the staffs of Houghton Library, Harvard University, and the Princeton University Library's Department of Rare Books and Special Collections provided support leading to the digitization and markup of Melville's marginalia to Shakespeare, Milton, and Homer. Jessica Ewing helped to coordinate markup for Melville's marginalia to all three authors represented in the cluster. Along with contributing to the Shakespeare analysis, Elisa Barney Smith supplied resources in MatLab from which the entire cluster has benefited. Much of the text analysis work informing the cluster was galvanized at the 2017 National Endowment for the Humanities Advanced Institute in the Digital Humanities, "Make YOUR edition: models and methods of digital textual scholarship" (Pittsburgh, 2017). Special thanks go to the conveners—in particular David Birnbaum, Ronald Dekker, Tara Andrews, and Leif-Jöran Olsson—for their training modules and their encouraging textual scholars to examine their edition data in new ways with computational methods. Much gratitude also goes to Mike Kestemont, who provided a helpful tutorial on R and the "stylo" package at the NEH Institute. An earlier version of this work received helpful feedback from the Works in Progress Seminar at the Institute of English Studies, University of London.

[1] Recovery of the erased words, with conjectural readings given in brackets, was achieved collaboratively by Dennis C. Marnon, Peter Norberg, and Steven Olsen-Smith.

[2] For another example of a multi-faceted approach that combines distant and close readings, see Janicke et al. Martin Eve has a forthcoming book with Stanford University Press, *Close Reading with Computers*, which will explore the idea of using digital approaches to convey depth and distance in reading practices for smaller sets of texts. For a variety of perspectives on Franco Moretti's idea of "distant reading," see also the May 2017 special section of *PMLA*, "On Franco Moretti's Distant Reading."

[3] For example, Anthony Kenny discusses the "mysterious veneration" that some literary scholars have for single and rare word occurrences, when "the rate of occurrence of a dull common word in a text may be a much more significant feature" (67–68). Similarly, Burrows, in *Computation into Criticism*, bases his analysis on the 30 most common words in Jane Austen's novels, with less attention to unique words.

[4] The GitHub repository can be found at <https://github.com/monline/leviathan-20.2.git>.

[5] Hammond suggests that one of the failures of distant reading comes from its lack of discoveries and tendency to over-validate its tools. When unique results are generated, they often cannot be verified, in his estimation.

Works Cited

Baayen, R. H. *Analyzing Linguistic Data: A practical introduction to statistics*. Cambridge: Cambridge UP, 2008.

Bond, Sarah E., Hoyt Long, and Ted Underwood. "'Digital' is Not the Opposite of 'Humanities.'" *Chronicle of Higher Education*, 1 Nov. 2017. Web. <https://www-chronicle-com.libproxy.boisestate.edu/article/Digital-Is-Not-the/241634>.

Brennan, Timothy. "The Digital Humanities Bust." *Chronicle of Higher Education*, 15 October 2017. Web. <https://www.chronicle.com/article/The-Digital-Humanities-Bust/241424>.

Burrows, J. F. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Oxford UP, 1987.

———. "Textual Analysis." In *A Companion to Digital Humanities*. Ed. Susan Schreibman, Ray Siemens, John Unsworth. *A Companion to Digital Humanities*. Oxford: Blackwell, 2004. Web. <http://www.digitalhumanities.org/companion/>.

Eder, M., Rybicki, J. and Kestemont, M. "Stylometry with R: a package for computational text Analysis." *R Journal* 8.1 (2016): 107–21.

Hammond, Adam. "The double bind of validation: distant reading and the digital humanities' 'trough of disillusionment.'" *Literature Compass* 14.8 (August 2017): e12402. Web.

Jänicke, S., et al. "Visual Text Analysis in Digital Humanities." *Computer Graphics Forum* 36.6 (2017): 226–50.

Jockers, Matthew. *Macroanalysis*. Urbana-Champaign and Chicago: U of Illinois P, 2013.

———. *Text Analysis with R for Students of Literature*. New York: Springer, 2014.

Kenny, Anthony. *The Computation of Style*. Oxford: Pergamon, 1982.

Mandell, Laura. "Experiencing the Bust." Blog post, 30 Oct. 2017. Web. <http://idhmc.tamu.edu/node/191>.

Melville, Herman. "Hawthorne and His Mosses." In *Piazza Tales and Other Prose Pieces, 1839–1860*. Ed. Harrison Hayford, Alma A. MacDougall, G. Thomas Tanselle et al. Evanston and Chicago: Northwestern UP and The Newberry Library, 1987. 239–53.

———. "Melville's Marginalia in the Dramatic Works of William Shakespeare." *Melville's Marginalia Online*. Boise State U. Ed. Steven Olsen-Smith, Peter Norberg, and Dennis C. Marnon. Web. 1 Jan. 2018.

Shakespeare, William. *The Dramatic Works of William Shakespeare*. 7 vols. Boston: Hilliard, Gray, 1837.

Silge, Julia and David Robinson. *Text Mining with R: A Tidy Approach*. Sebastopol, CA: O'Reilly Media, Inc, 2017. Web. <https://www.tidytextmining.com/>. 5 January 2018.

"On Franco Moretti's Distant Reading." *PMLA* 132.3 (May 2017): 613–89.

Turpin, Zachary. "Melville, Mathematics, and Platonic Idealism." *Leviathan: A Journal of Melville Studies* 17.2 (June 2015): 18–34.