Digital Text Analysis of Herman Melville's Marginalia in Shakespeare [A Progress Report]

By Christopher Ohge September 13, 2018

In Uncategorized No Comments

(Click here to download all of the slides as a PDF file.)

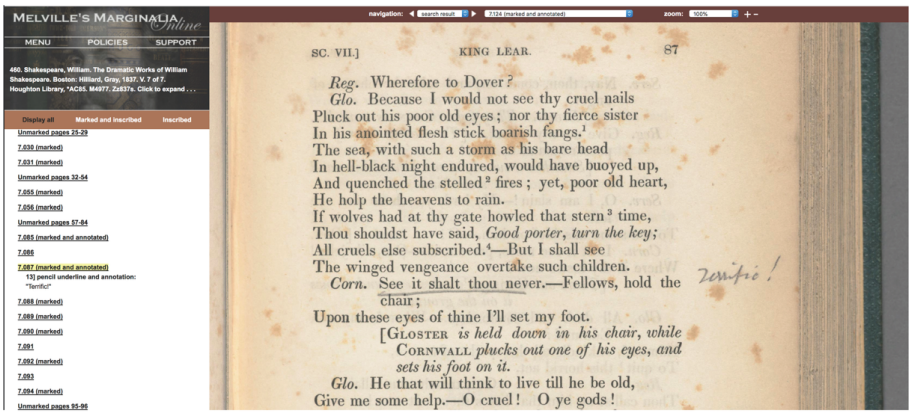
"He had the tradition in him, deep, in his brain, his words, the salt beat of his blood. He had the sea of himself in a vigorous,

stricken way… It enabled him to draw up from Shakespeare… History was ritual and repetition when Melville's imagination was at its own proper beat."

––Charles Olson, *Call Me Ishmael* (1947)

Melville was a keen appreciator of Shakespeare––this is not a groundbreaking revelation to many. Yet few scholars have closely examined Melville's marginalia in Shakespeare to reveal that influence. In his copy of *King Lear*, for example, Melville noted many of the concise passages involving Lear's tragic double-bind in the play. In the exchange that precipitates Gloster's blinding, when Gloster declares he will live to see vengeance delivered upon Lear's daughters, Cornwall responds, "See it shalt thou never!" Melville underlined the diabolical wit of the rebuttal and wrote in the margin, "Terrific!"

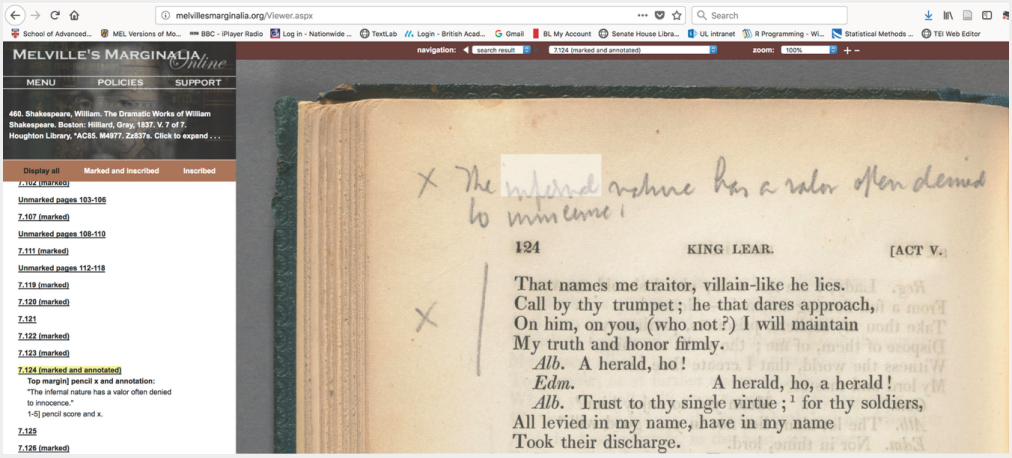
(Click on the images to expand.)



Terrific, as in terrifying (think of "The Serpent … with brazen Eyes

| And hairie Main terrific" in Book VII of *Paradise Lost*). And terrifically true to Melville.

In the final scene of *King Lear*, when Edmund responds to Albany's challenge by exclaiming "I will maintain | My truth and honor firmly," Melville wrote at the top: "The infernal nature has a valor often denied to innocence."

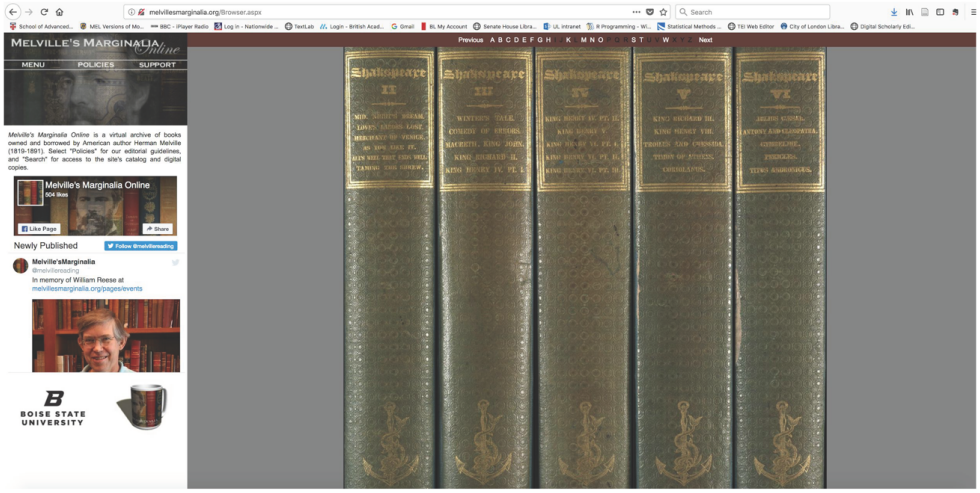


Consider that annotation in relation to the innocent nature of Starbuck in *Moby-Dick*: "That immaculate manliness we feel within ourselves, so far within us, that it remains intact though all the outer character seem gone; bleeds with keenest anguish at the undraped spectacle of a *valor*-ruined man" (Chapter 26, "Knights and Squires"). And Ahab's valor does not seem to be isolated within himself but rather partakes in a natural will of malice, such that his crew "seemed specially picked and packed by some *infernal* fatality to help him to his monomaniac revenge" (Chapter 41, "Moby Dick"). "The *infernal* nature has a *valor* often denied to innocence."

I show these examples from *King Lear* to illustrate the

importance of the evidence of Melville's close engagement with Shakespeare, on the one hand, as well as the tendency to focus on annotation in studies of authorial reading. Markings such as underlinings, marginal scores, and checkmarks are just as important and can also reveal patterns of reading that affected the Melville's thinking.

Melville's Marginalia Online <melvillesmarginalia.org>



Melville's marginalia in his 1837 American edition of the Hilliard, Gray *Dramatic Works of William Shakespeare* offer a strong case for digital text analysis among books that survive from his library. He marked thirty-one plays in the seven-volume set, comprising 681 distinct passages with marginalia that can be attributed to Melville. Previous attempts by scholars to count the marginalia were significantly off, but working with Steven Olsen-Smith and a team of contributors, I have applied computational approaches to reading the data of the marginalia. What this shows is that text mining approaches, while often used for large data sets, can also effectively aid close reading and facilitate new discoveries. This kind of text analysis may be even more

important for marginalia, which is fragmentary by nature. It is a curated selection of words within a coherent text.

Interpretive stakes are high for analysing Melville's marginalia to Shakespeare's plays. Aside from his comments on his friend and fellow author Nathaniel Hawthorne in "Hawthorne and His Mosses," Melville's pronouncements on Shakespeare in the same essay constitute his most detailed assessment of a writer whose works survive, heavily marked and annotated, from his library. Their main purport boils down to the following extraordinary passage:

> "But it is those deep far-away things in him; those occasional flashings-forth of the intuitive Truth in him; those short, quick probings at the very axis of reality;—these are the things that make Shakespeare, Shakespeare. Through the mouths of the dark characters of Hamlet, Timon, Lear, and Iago, he craftily says, or sometimes insinuates the things, which we feel to be so terrifically true, that it were all but madness for any good man, in his own proper character, to utter, or

> even hint of them."

Shakespeare's genius hinges on interconnected notions of rhetoric, sentiment, and reception. His most profound disclosures, with their philosophically bleak implications, are made in few words, whether more or less directly but stealthily and by insinuation. Their potentially baneful effects on readers––who are intellectually and temperamentally unprepared for them––necessitate that craftiness.

Melville's Marginalia Online is one of the most advanced digital projects devoted to an author's personal library and marginalia. In addition to the data entry for the front end of the digital archive, staff members of the project have been also marking up the digitised books by Melville using coordinate-capture XML markup, which allows for word searching that also highlights the results on the facsimile image of the book page.

See, for example Melville's first marking in *The Tempest*:

```
<div id="2" x="277" y="2415" group="1" width="
<w x="416">That</w>
<w x="526">this</w>
<w x="653">lives</w>
<w x="726">in</w>
<w x="815">thy</w>
<w x="1023">mind?</w>
<w x="1197">What</w>
<w x="1344">seest</w>
<w x="1469">thou</w>
<w x="1574">else</w>
<div id="3" x="277" y="2479" group="1" width="
<w x="353">In</w>
<w x="446">the</w>
<w x="580">dark</w>
<w x="836">backward</w>
<w x="943">and</w>
<w x="1124">abysm</w>
<w x="1192">of</w>
<w x="1345">time?</w>
</div>
</div>
```

Each marking is encoded as a `<div>` element which comes with several attributes identifying aspects such as the play to which it belongs, the play's mode, and other relevant information that aid the text analysis. Each word is marked with a word element (`<w>`) and a coordinate attribute that points to its place on the page facsimile. The example also features an embedded marking (an underline) within the checkmarked passage.

As a result of this XML markup, a user can now undertake word searches on the site. Here I have entered "fear," which as I show later, is a prominent negative-sentiment word among Melville's markings.

Melville marked the word "fear" in several of his books, including his other major source for *Moby-Dick*, Thomas Beale's *Natural History of the Sperm Whale*. But there the word appears 17 times in the Shakespeare marginalia.

By clicking on the Shakespeare results one can see more details about the words––the first result features two instances of "fear" as well as a corresponding annotation.

One can click on that particular result to see the book page with the search results highlighted therein. Here Melville is showing his intertextual prowess, cross-referencing the death-as-sleep metaphor in *Measure for Measure* with another instance of it in *The Tempest*. These individual and nuanced explorations of Melville's marginalia are a huge boon to researchers, but we are now using text analysis and data mining to reveal more about the total corpus of these markings.

How important was Shakespeare to Melville's style? As an experiment, I ran a stylometry calculation using the "stylo" package in R.

Stylo's accuracy shows in this visualisation, 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.

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"––are related to perception. These discoveries are catalysts for new analytical directions.

Why might Shakespeare be closer than Homer and Milton? Stylo also reveals that Melville, like Shakespeare, was drawn more to the first person (unlike Homer and Milton). This R code (shown in RStudio) bolsters this conclusion:

A table of the most frequent bigrams in the Shakespeare marginalia, highlighted in red, shows a heavy prevalence of first-person constructions––five out of the top ten, in fact. No other pronouns appear.

The XML encoding of each marking on the word-level facilitates computational analyses of the markings.

With an XSLT transformation, we produced word counts for each play.

One can quickly learn a lot––and pose some questions––from these results. There are some surprises: Melville marked more words in the comedies than in the tragedies; that among the comedies, he marked the most words in *Measure for Measure* (well, that might not be so surprising, because it is a dark comedy); that the tragedy with the most markings is *Antony and Cleopatra*; that he marked more words in *Henry VIII* than he did in *Hamlet* or *King Lear*. We can already see how far we have come since I showed you the two intriguing annotations in *King Lear*––a play which actually represents a small percentage of his notes in Shakespeare.

Given the apparent differences between the comedies and tragedies, we realised the necessity to calculate the word

counts-per-marking, as well as the average word count per marking, in each play mode.

It turns out that Melville marked much shorter passages in the tragedies than in the comedies––an average difference of about 10 words.

R code adapted from Matthew Jockers's *Text Analysis with R for Students of Literature* can also calculate other linguistic features. Here is a graph of lexical uniqueness (calculating hapax legomena, or words that only occur once in a corpus).

Again, the markings in the tragedies have the highest lexical variety––that is, the highest percentage of unique words. This can also be read as a general index of briefness in passages and sections. Now let's consider the low average word counts and high lexical variety in the tragedies in light of Melville's quote about Shakespeare: "But it is those deep far-away things in him; those occasional flashings-forth of the intuitive Truth in him; those short, quick probings at the very axis of reality;—these are the things that make Shakespeare, Shakespeare."

Short, quick probings––as reflected in his own notes to the text. Construed, then, within the framework of esoteric expression he attributed to Shakespeare, Melville's preoccupation with the bleakness of worldly and human conditions in his marginalia to the *Dramatic Works* corresponds with the views he expressed in "Hawthorne and His Mosses."

Melville's mentioning the four primary dark characters moved us to test the lexical uniqueness in the plays in which they appear.

Indicators of brief marked passages carry heightened prospects of significance; for it is in such marginalia that we may expect to encounter what Melville described as Shakespeare's "short quick probings at the very axis of reality". The passages with maxed values offer a number of different candidates for the sorts of disclosures Melville had in mind, including "Virtue itself of vice must pardon beg" (*Hamlet*) and "Truth's a dog that must to kennel" (*King Lear*). Overall this graph reveals more nuanced information than can be gleaned from the word count graph: the markings in *King Lear* contain almost three times more lexically

unique marked passages than *Hamlet*. It would be a mistake, therefore, to deduce from the marked-word counts that Melville engaged more extensively with *Hamlet* than with *King Lear*. Instead, he engaged with each differently, marking fewer but longer passages in *Hamlet*, and a larger number of shorter passages in *King Lear*.

The two markings in *Othello* also have quite divergent lexical values.

Low values can call attention to passages marked by Melville for their rhetorical qualities along with their purported sense, which is the case for the second of these passages but less so for the first. Here Melville's interest was focused primarily on the idea of Iago's dark utterance of incisive profundity. But in the second passage, Melville's attention to wordplay, as well as to sense, bespeaks a different but no less significant dimension of the

verbal features that moved him to apply pencil to paper.

If we start to investigate the words themselves with the aid of computation, the results suggest new avenues for understanding Melville's reading of Shakespeare. The three most common substantive words that Melville marked were "man", "world", and "love".

As this table shows, the high-frequency terms follow an interesting trajectory, showing that "love" appears proportionally the most in comedies, "world" in the tragedies, and "man" in the histories.

Moreover, a wordcloud of word frequencies in the comedies shows a lot of what might be considered humanistic terms, whereas in the tragedies there is a cluster of "world" in tension

with “man” and “good”.

In addition to these helpful word frequency results, I also created tables of all the markings with another XSLT script.

With that we are now able to provide HTML tables of each marking with its associated bibliographic reference as well as its word count.

Users of these tables for the recent special issue of *Leviathan* (June 2018) have already found it very efficient to have all the markings in one searchable table with their associated metadata.

The table is also sortable by word count, which is particularly important for gauging Melville's attention to brevity.

Or prolixity.

(To access the full table of markings, go to christopherohge.com/460-word-count-per-marking.html.)

These fairly rudimentary calculations already provide a good amount of research questions, but there are still many other ways to investigate this fairly small data set of marked words for close reading. One way of doing that is with sentiment analysis––which is particularly relevant to an author who,

according to Melville, was esoterically dark.

Drawing on Julia Silge and David Robinson's "tidytext" package in R, sentiment analysis shows the frequency of positive and negative words in a given data set. Melville clearly marked a much greater proportion of negative words in Shakespeare relative the whole texts, so he was sincere in his estimation of what he called Shakespeare's blackness.

The sentiment data also show that Melville marked a greater net number of negative words, and 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. A bar graph 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.

Recall at the beginning I looked at "fear," which is fourth from the top, and "death" ranks second from the top. But the positive words raise some questions as well––can "good" really be positive in Shakespeare, or "love" (also one of the highest frequency terms overall)? How does the word "great" function in context?[2] What can be inferred from the high frequencies of negative terms and the concentrated appearance of select positive terms?

To complement this data I produced a tidytext tibble (which is a dataframe that organises each variable in a column and each observation in a row. Each type of observational unit is a table––for us that boils down to one token per row with various attributes). We have organized the data such that each marking

observation has a corresponding play title. First I created a table of bigrams without stopwords to look for some more clues.

In the top ten results, already two interesting ones come out: "peace peace" and "hate thee"––both of which indicate trouble.

Next I created a trigram table: the presence of fewer trigram results reflects the small corpus size, so these might seem less relevant.

Yet it is still intriguing to notice the implication of tenuous personal relationships in the substantive trigrams, especially the ones highlighted in red. Testing trigrams is meant to achieve a sense of linguistic relations––subject, object, verb constructions can be very informative. In this case, however, what does the lack of repeatable trigrams show, other than the fact that this is a small corpus? It reinforces the previous results showing that Melville was generally not attending to repetition, that he was paying more attention to distinct utterances and ideas rather than rhetoric.

The following tibble shows the TF-IDF results of bigrams: standing for term frequency–inverse document frequency, TF-IDF is a numerical statistic that attempts to reflect the

importance of a word or group of words.

TF-IDF is often weighted by the number of occurrences in a document and its appearances throughout the corpus, but here we can see the striking results within one document of Melville's markings. Particularly with bigrams, the results suggest the importance of pessimistic pairings that are quite unique in the corpus: "life cancels," in *Henry IV* (a play that scholars may have overlooked as an influence on Melville), but also the group from *Othello*: "blood burns", "dangerous conceits", "nature's poisons", "poison natures".

A larger table of TF-IDF results also suggests some important bigrams in *Richard III*, with "darkness true" and "meaner

creatures".

In *Romeo and Juliet*, "blood stirring" and "mad blood"; and in the *Taming of the Shrew*, "bitter word".

And these bigrams are bespeaking bitterness, indeed. Of course they are not meant to suggest that the TF-IDF bigrams are the most important pairings among Melville's markings, but they are guideposts to new avenues of his reading. They also emphasise the weight of negativity (with more context) in the markings. However, they are statistically provocative within the document, and are worth investigating further.

Having generated these bigram tables, I can also undertake other sentiment analyses in which the code finds pairings of the most frequent words preceded by "not". This provides more context to the unigram sentiment results I showed earlier.

Given the dark implications of the markings, notice too that the highest frequency word to be preceded by "not" is the word "good." Tinkering the R code further to inner-join a vector of negation words with the sentiment calculation, I also produced a graph of sentiment words preceded by negations.

And again, the word "good" comes out on top, but some other interesting results appear: "satisfied", "pleasure", and "true". Even the negation of a seemingly negative word such as "pity" provides no less comfort.

Returning to the negative sentiment unigrams, a wordcloud of all the negative sentiment terms weighted by frequency can guide further exploration, complemented by the overall word frequencies.

Making sense of all these analyses requires some critical thinking and close readings of these marked words and ideas in relation to Melville's own works. Given that we know Melville was reading Shakespeare closely while writing *Moby-Dick*, we focussed our attention to the play *Henry VIII* (a play which we might have ignored given its lack of notoriety but which we now investigate because it was the third-highest marked play). For example, he scores a speech by Cardinal Wolsey, which contains several terms duplicated or approximated in the negative wordcloud, such as "fear," "weak," and "malice":

> "We must not stint
> Our necessary actions, in the fear

To cope malicious censurers; which ever,
As ravenous fishes, do a vessel follow
That is new trimmed; but benefit no further
Than vainly longing. What we oft do best,
By sick interpreters, once weak ones, is
Not ours, or not allowed; what worst, as oft,
Hitting a grosser quality, is cried up
For our best act. If we shall stand still,
In fear our motion will be mocked or carped at…"

These passages, never before cited as a potential source for *Moby-Dick*, presage the language of Ahab's tragic striving, and his malicious and monomaniacal quest.

Ahab's obsession for vengeance, which the narrator Ishmael compares to the obstinacy of a "thunder-cloven old oak" (125), also resonates with Melville's attention to the dialogue in *Measure for Measure* featuring Isabel's comments about misappropriated strength and the tyranny of hubris: "And he, that suffers. O, it is excellent / To have a giant's strength; but it is tyrannous / To use it like a giant" (1: 358). Isabel's next remark on the ludicrous pride of "man" (quoted earlier) is preceded by a suggestive analogy to the natural and phenomenal world:

> "Thou rather, with thy sharp and sulphurous bolt,
> Split'st the unwedgeable and gnarled oak,
> Than the soft myrtle:––But man, proud man!
> Dressed in a little brief authority,—
> Most ignorant of what he's most assured,
> His glassy essence,—like an angry ape,
> Plays such fantastic tricks before high Heaven,
> As make the angels weep; who, with our spleens,
> Would all themselves laugh mortal."

Involving the high-frequency term "man" and the prominent sentiment-result "heaven," as well as the sentiment graph data relating to pride, anger, strength, suffering, and death, the page of *Dramatic Works* containing Isabel's laments also features her speculation, "could great men thunder," and repeats "thunder" multiple times. Melville also marked Cleopatra's comparing Antony to "rattling thunder."

As you can see above, the image of a thunderstruck tree and the presence of outsized strength and hubris all figure in the simile at the end of chapter 119 of *Moby-Dick*: "As in the hurricane that sweeps the plain, men fly the neighborhood of some lone, gigantic elm, whose very height and strength but render it so much the more unsafe, because so much the more a mark for thunderbolts; so at those last words of Ahab's many of the mariners did run from him in a terror of dismay." The following examples from *Henry VIII*, *Measure for Measure*, and *Antony and Cleopatra* all illustrate the confluence of influences from plays across modes that also happen to have been among the plays in which he left the most markings. It is difficult to imagine Ahab's character coming into being with such force––and nuance––if Melville had not studied Shakespeare's plays.

Computational approaches to Melville's marginalia allow

readers to calculate word counts and frequencies, word variety, topic clusterings, and sentiment associations. Complemented with informed acts of careful reading and source elucidation, these text analyses reveal Melville constructing new paths in his own writing from his experiences of reading Shakespeare. By using distant reading strategies with the marginalia, in their own right and in the service of close reading, we arrive more informed than ever at the "very axis" of their genius.

The methods demonstrated in this talk were inspired by Matthew Jockers's *Text Analysis with R for Students of Literature* and Julia Silge and David Robinson's *Text Mining with R*. The XML files, XSLT, and R code used by the MMO team can be accessed on GitHub at https://github.com/monline.

---

1. Parts of this talk were adapted from a lengthier essay that was published as part of a special issue of *Leviathan: A Journal of Melville Studies* 20.2 (June 2018) devoted to digital text analysis of Melville's marginalia, but this talk also revealed some new techniques and visualisations that did not make it into the publication (which can be accessed at http://muse.jhu.edu/issue/38709). It is important to stress that despite the illusion of completeness that a publication suggests, the text analysis of Melville's marginalia is an organic process. We are still refining and improving our methods and learning more about

ways to enhance our understanding of Melville's reading. ↩

2. In our *Leviathan* piece, we show that the word "great" is not always a positive word. Sometimes it is an enhancer of a negative word; other times it is just a flavoring word. This kind of word demonstrates the necessity for critical thinking of the data. ↩

Leave a Reply

Your email address will not be published. Required fields are marked *

comment *

name *

email *

website

post comment