

## The Spenser Review

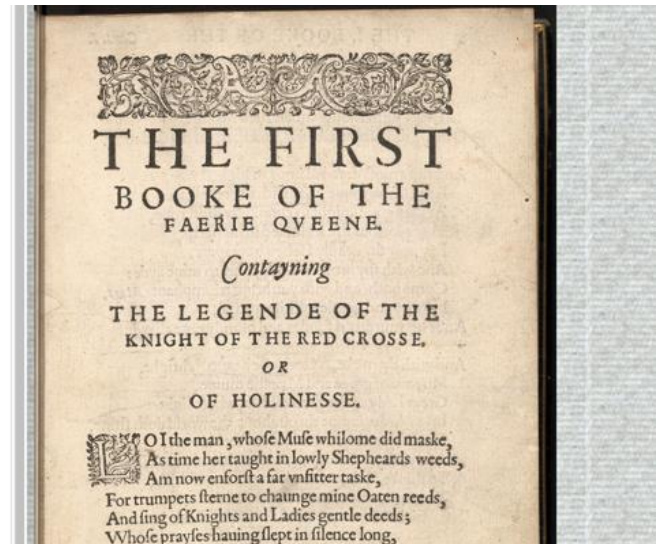
The faerie queene Disposed into twelue bookes, fashioning XII. morall...

THE FIRST BOOKE OF  
THE FAERIE QVEENE. Contayning  
THE LEGENDE OF THE KNIGHT  
OF THE RED CROSSE  
OR OF  
HOLINESSE

[Image 002-b]

id: A12778-002-b-0300  
lemma: whilom  
reg: whilom  
pos: av

LOI the man, whose Muse whilome did maske,  
As time her taught in lowly Shepheards weeds,  
Am now enforst a far vnfitter taske,  
For trumpets sterne to change mine Oaten reeds,  
And sing of Knights and Ladies gentle deeds;  
Whose prayes hauing slept in silence long,  
Me, all too meane, the sacred Muse areeds  
To blazon broad amongst her learned throng:  
Fierce warres and faithfull loues shall moralize my song.



## The Lab and the Library: An Introduction to the EarlyPrint Project<sup>1</sup>

By Craig A. Berry (Independent Scholar)

“LOI [*sic*] the man, whose Muse whilome did maske . . .” So begins the proem to Book I of the 1596 *Faerie Queene* in the transcription produced by the Text Creation Partnership (TCP) from the Early English Books Online (EEBO) page images.<sup>2</sup> Spenserians may be annoyed or amused (or both) by the mistranscription of the opening words “LO” and “I” as the single word “LOI,” as if there were some poet named Loi who is bracing himself and his muse for the task ahead. Even if the line were not so famous, we would realize that “loi” is not a word in English, that “I” must be a separate word for the verb “Am” that opens line 3 to make any sense (“Am now enforst a far vnfitter taske”), and that the meter does not work when two words are collapsed into one. In echoes of this line by hands far less steady on the metrical rudder than Spenser’s, as well as in the 1590 *Faerie Queene*, the EEBO-TCP transcription contains the exact same error.<sup>3</sup> We can

<sup>1</sup> The project home page may be found at <https://earlyprint.org>. In the interest of full disclosure, I am a member of the project team, and this essay is not a review, but more of a tour of the premises by someone who lives there.

<sup>2</sup> Edmund Spenser, *The faerie queene Disposed into twelue bookes, fashioning XII. morall vertues* (London: William Ponsonbie, 1596; Ann Arbor: Text Creation Partnership, 2008), I.proem.1, <https://quod.lib.umich.edu/e/eebo/A12778.0001.001/1:3.1?rgn=div2;view=fulltext>.

<sup>3</sup> “LOI [*sic*] the man whose Muse did lately forage” in John Taylor, *All the vvorkes of Iohn Taylor the water-poet* (London: James Bolter, 1630; Ann Arbor: Text Creation Partnership, 2005), 61, <https://quod.lib.umich.edu/e/eebo/A13415.0001.001/1:30.4?rgn=div2;view=fulltext>; “LOI [*sic*] the Man, whose Muse mus’d on Plantations” in Samuel Purchas, *Purchas his pilgrimes, part 4 In fiue bookes* (London: Henrie Fetherstone, 1625; Ann Arbor: Text Creation Partnership, 2004) 1891, <https://quod.lib.umich.edu/e/eebo/A71306.0001.001/1:12.1.1?rgn=div3;view=fulltext>; and Spenser, *The faerie queene* (London: William Ponsonbie, 1590; Ann Arbor: Text Creation Partnership, 2002), <https://quod.lib.umich.edu/e/eebo/A12777.0001.001/1:3?rgn=div1;view=fulltext>.

## The Spenser Review

speculate on the reason for the error by looking at page images of the 1590 (fig. 1) and 1596 (fig. 2) editions:

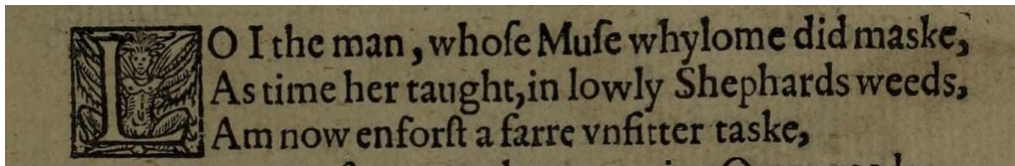


Fig. 1. 1590 *Faerie Queene*, Boston Public Library, PR2358.A1 1596a.<sup>4</sup>

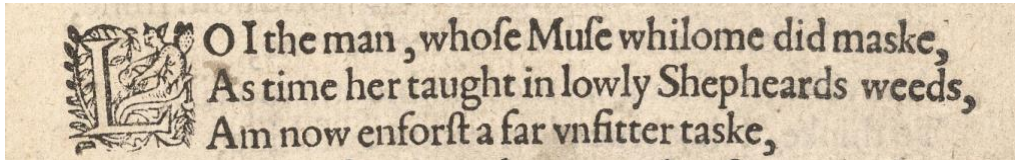


Fig. 2. 1596 *Faerie Queene*, STC Collection, The University of Pennsylvania, PR2358.A1 1596.<sup>5</sup>

In 1596, the “O” appears to be suspended mid-way between the “L” and the “I,” but in the 1590, the “O” snuggles a bit closer to the decorated initial capital and has more space after it, yet the transcriptions of both fail to recognize that the following “I” is a separate word. The TCP transcribers had instructions to record exactly what they saw on the page, but regarding spaces, the guidance provided to them notes: “In cases of doubt, it may be necessary to use the sense of the passage to dictate its spacing.”<sup>6</sup> It appears that either the expectation of an initial word in all capitals won out over the sense of the passage or that the sense was simply opaque to the transcribers, yielding a result that is clearly disappointing to early modernists even though such errors represent a proportionally small number of the transcribed words. But early modernists did not produce the TCP transcriptions, and the keyboardists who did so recorded approximately 1.5 *billion* words in digital form without any wear and tear on early modernist carpal tunnels. So perhaps we can mix some humility and gratitude with our disappointment, and better yet, we can do something about that disappointment by building on what we have been given.<sup>7</sup> The [EarlyPrint](https://earlyprint.org) project is about doing everything that can be done for and with this immense, flawed, yet invaluable digital record of the English print heritage.

The *EarlyPrint* project has been built over a period of several years with generous

<sup>4</sup> Accessed via *The EarlyPrint Library*, <https://texts.earlyprint.org/works/A12777.xml?page=002-b>.

<sup>5</sup> Accessed via *The EarlyPrint Library*, <https://texts.earlyprint.org/works/A12778.xml?page=002-b>.

<sup>6</sup> “Keying/Coding Specifications” (Ann Arbor: Text Creation Partnership, 2001), <https://textcreationpartnership.org/docs/dox/instruct3.html>.

<sup>7</sup> For more on what it is that we’ve inherited from the TCP, see “The TCP Texts and their Shortcomings,” <https://earlyprint.org/intros/about-ep-texts.html>.

## The Spenser Review

support from the Andrew W. Mellon Foundation, the American Council of Learned Societies, and several contributing institutions, most notably Northwestern University and Washington University in St. Louis. *EarlyPrint* consists of a website with two main divisions, respectively called the *EarlyPrint Library* and the *EarlyPrint Lab*, but the metaphor of campus buildings gets us only so far in understanding what the site offers. The *Library* is indeed a place to find books and read them, and the *Lab* facilitates research and experimentation, but the containers involved are neither bookshelves nor test tubes but rather tags in eXtensible Markup Language (XML). These tags contain document structure (such as stanzas, lines of verse, or chapter divisions) as well as words and their attributes and provide the underpinnings of all of the many features of both the *Library* and the *Lab*. The word-level tagging in particular distinguishes the TCP-derived texts available in *EarlyPrint* from the original TCP transcriptions, as we can see by comparing that famous opening line in its TCP version with its enhanced *EarlyPrint* version.<sup>8</sup> In TCP, the line is simply wrapped in an “l” element, indicating a line of verse, and a special notation to capture the decorated initial has been added as shown in figure 3:

```
<l><seg rend="decorInit">L</seg>OI the man, whofe Mufe whilome did maske,</l>
```

Fig. 3. TCP transcription of 1596 *Faerie Queene* I.proem.1, line 1.

Consider the same line in *EarlyPrint*, where figure 4 shows that the long “s” has been normalized and each word has been assigned a unique identifier and lives within its own “w” element:

```
<l xml:id="A12778-e200">
  <w lemma="n/a" pos="ffr" rend="decorinit" xml:id="A12778-002-b-0240">LOI</w>
  <w lemma="the" pos="d" xml:id="A12778-002-b-0250">the</w>
  <w lemma="man" pos="n1" xml:id="A12778-002-b-0260">man</w>
  <pc xml:id="A12778-002-b-0270">,</pc>
  <w lemma="who" pos="crq" xml:id="A12778-002-b-0280">whose</w>
  <w lemma="muse" pos="n1" xml:id="A12778-002-b-0290">Muse</w>
  <w lemma="whilom" pos="av" reg="whilom" xml:id="A12778-002-b-0300">whilome</w>
  <w lemma="do" pos="vvd" xml:id="A12778-002-b-0310">did</w>
  <w lemma="mask" pos="n1" reg="mask" xml:id="A12778-002-b-0320">maske</w>
  <pc xml:id="A12778-002-b-0330">,</pc>
</l>
```

Fig. 4. *EarlyPrint* encoding of 1596 *Faerie Queene* I.proem.1, line 1.

---

<sup>8</sup> On the editorial practices and encoding changes involved, see “From TCP to *EarlyPrint*: the conversion process,” [https://earlyprint.org/intros/tcp\\_to\\_earlyprint.html](https://earlyprint.org/intros/tcp_to_earlyprint.html).

## The Spenser Review

The identifier in the “xml:id” attribute makes each of the corpus’s 1.5 billion words independently addressable, which in turn allows precise identification of search results (and words that need fixing, such as “LOI”). When a standard spelling is different from the original spelling, such as “mask” for “maske” or “whilom” for “whilome,” the standard, or regular, spelling is recorded in a “reg” attribute, which allows spelling variances to be leveled out in searches and also provides the ability to switch dynamically between old and standard spellings in a reading display. The lemma, or dictionary head word, similarly allows the leveling out of inflection during search or analysis. The part of speech recorded in the “pos” attribute identifies the grammatical function of the word using the NUPOS tag set, a linguistic classification system created by Martin Mueller, one of the EarlyPrint principal investigators, and developed specifically to handle the eccentricities of Early Modern English. All of these attributes together make the text and the entire corpus into a network of words whose digital representation allows it to be searched and analyzed on a variety of different axes. The NUPOS tags are created by *MorphAdorner*, a program developed by the late Philip R. Burns that “tokenizes” the text into words and assigns such “morphological adornments” as lemma and part of speech to each word using frequency data and a variety of heuristics.<sup>9</sup> *MorphAdorner* is quite accurate most of the time, but it can miss the mark, especially if given bad input: that “LOI” in our opening line above has been identified as a foreign word in French is a pretty good guess for a computer to make given that there is no reasonable way to construe this as a single word in English. Of course, there is no French word here, and in this case a human-produced error has led to a computer-produced one. Better input leads to better output, and the example illustrates the importance of continuous improvement of the transcriptions in order to facilitate both the readability of the texts and their automated processing. However, there are more texts with a low density of errors than there are with a high density, and with a glass of correctly-tagged words that is more than 95% full we can now proceed to explore the features of the *Library* and *Lab* that build on these word-level data.

The *Library* component of *EarlyPrint* has many of the features of its physical namesake but aims to bridge the gap between early modern book culture and twenty-first-

---

<sup>9</sup> Philip R. Burns, *MorphAdorner*, <http://morphadorner.northwestern.edu/morphadorner/>. For a review of *MorphAdorner* in these pages, see Anupam Basu, “MorphAdorner v2.0: From Access to Analysis,” *Spenser Review* 44.1.8 (2014), <http://www.english.cam.ac.uk/spenseronline/review/volume-44/441/digital-projects/morphadorner-v20-from-access-to-analysis>.

## The Spenser Review

century book culture. The list of roughly 65,000 books functions more or less as a card catalog, but instead of rows of alphabetized drawers there are over a dozen filters to facilitate locating the book or set of books of interest. Criteria include basic search categories such as author, title, and year, as well as more project-specific criteria such as who has previously contributed corrections to the transcription (a “curator”), and what the text’s “grade” is, where grade measures the density of known defects in the transcription. For example, a “B” text has one to ten known defects per 10,000 words, and an “A” has no known defects.<sup>10</sup> In addition to the text filter, which operates on metadata, there is also a simple text search that operates on words and allows searching by any of the word attributes mentioned above. Figure 5 shows the [Library main page](#), with the text filter on the left, the list of texts in the middle, and word search on the right.

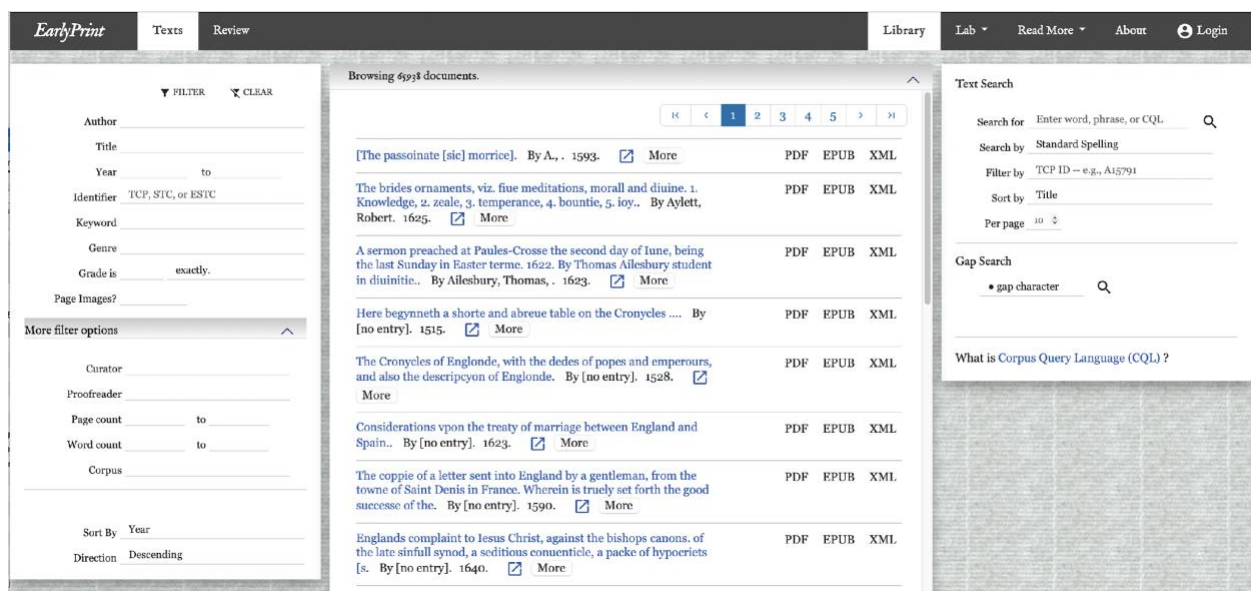


Fig. 5. Browsing texts in *The EarlyPrint Library*.

Selecting a text opens a reading pane with a display that can be paged through, but the codex metaphor is merely the foundation, with many other capabilities layered on top.<sup>11</sup> Hovering over a word reveals a tooltip with the word’s linguistic attributes, while an options menu allows switching back and forth between standard and original spellings. A “Downloads” menu allows the book to be downloaded in ePub or PDF formats for

<sup>10</sup> For a fuller explanation of the grading system, see “Curation and Quality Assurance,” <https://earlyprint.org/posts/curation-and-quality-assurance.html>.

<sup>11</sup> For more on the basics of navigation, see “How To Browse, Search, and Display,” [https://earlyprint.org/how-to/browse\\_search\\_display.html](https://earlyprint.org/how-to/browse_search_display.html).

## The Spenser Review

reading on a tablet or computer, but also makes available the raw XML that you may wish to use as a starting point for producing your own edition or as the data set for your own analysis. An “EEBO” link in the navigation bar will, if your institution subscribes, bring up the same page in EEBO. Figure 6 shows the opening of the 1596 *Faerie Queene* in the *Library* reading view with a word attribute tooltip displayed above “whilome” in line 1

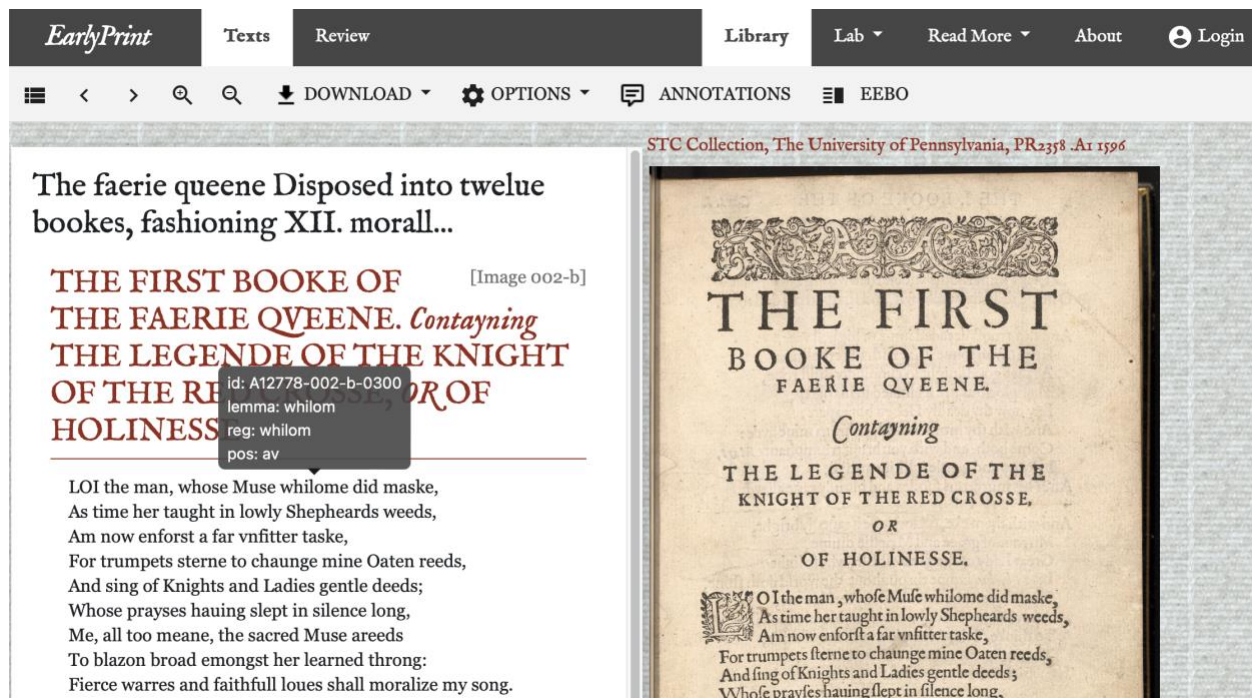


Fig. 6. 1596 *Faerie Queene* opening lines in *The EarlyPrint Library*.

The text in the reading pane is a writerly one that encourages the active participation of the reader, though not in a sense that has anything to do with postmodern hermeneutics: the user can propose corrections to gaps and transcription errors, either those encountered along the way while reading, or those intentionally hunted down using the “gap search” capability.<sup>12</sup>

The ability to correct errors in the transcriptions is the most distinctive (and possibly the most valuable) feature of the *Library*. The practice of crowd-sourced “collaborative curation,” as it is called, takes some getting used to for those of us accustomed to accepting whatever libraries and publishers provide.<sup>13</sup> It better resembles

<sup>12</sup> To get started correcting, see the tutorial, “How to Correct Gaps,” [https://earlyprint.org/how-to/howto\\_correct\\_gaps.html](https://earlyprint.org/how-to/howto_correct_gaps.html).

<sup>13</sup> We have discussed data curation at more length in Martin Mueller, Philip R. Burns, and Craig A. Berry, “Collaborative Curation and Exploration of the EEBO-TCP Corpus,” in *Early Modern Studies after the Digital Turned*, ed. Laura Estill, Diane K. Jakacki, and Michael Ullyot (Tempe: Arizona Center for Medieval and Renaissance Studies, 2016), 145-65. See also Mueller’s essay in these pages, “The EEBO-TCP Phase I

## The Spenser Review

open-source software than traditional publishing, which is to say that anyone anywhere in the world with web access can propose fixes to textual defects. One must first create an account and log in, after which clicking on any word in the text display brings up an annotation pane in which revised word transcriptions may be entered. There are special facilities for emending wrongly joined words (such as “LOI”) or wrongly split words, but the simplest option is just to record a new word transcription that replaces a problematic one. In due course, that correction will be reviewed by the site editors and, if approved, incorporated into the text. Some tens of thousands of corrections have been made thus far, in many cases by undergraduates engaged in class projects or working as summer interns. There are about five million remaining words with known defects, so there is plenty of work left to be done.

Most of the known defects in the transcriptions arise from the fact that the transcribers were instructed not to guess when they encountered illegible or undecipherable page content. This leaves us with many cases of black dot characters (“•”), indicating untranscribed letters, or lozenge characters (“◇”) indicating untranscribed words. Some knowledge of early modern spelling and diction and some experience with deciphering the effects of over-inking, broken type, and other vagaries of early printed books and their perilous journey through time will make a noticeable difference in resolving these undecipherables via another look at the EEBO image. However, a modern, full-color page image in high resolution is almost always vastly superior to EEBO and offers the next best thing to the original book for fixing transcription problems. The *EarlyPrint Library* provides a way to show such images side-by-side with the transcription by taking advantage of the International Image Interoperability Framework (IIIF) to link to image sets at institutions around the world. The *Library* currently has close to a thousand matching image sets that show page images next to transcriptions in a format that is similar to a facing-page translation (visible above in Figure 6), and more such image sets will surely become available as rare book libraries continue to digitize their holdings and serve them up via IIIF. Most libraries offer their own interfaces for viewing the images, but by incorporating the images into its own interface, the *EarlyPrint Library*, via its coordinated set of search, display, and correction features with side-by-

## The Spenser Review

side transcriptions and images, simultaneously provides multiple avenues of mediation into the world of an early printed book.

The *EarlyPrint Lab* provides analytical tools for examining the linguistic, structural, and bibliographical characteristics of the enhanced TCP corpus. Its physical and spiritual roots lie in the Humanities Digital Workshop at Washington University in St. Louis where *EarlyPrint* co-principal investigator Joe Lowenstein is Director, and where a long-held penchant for collaboration and experimentation has coalesced with the *EarlyPrint* data to provide unique opportunities for early modernists. The sort of activities the *Lab* provides and enables have been famously (or infamously) called “distant reading” by Franco Moretti, a clever phrasing that allows him to land one more kick on the bruised torso of close reading by claiming to offer its opposite, and also implies that the new digital methodology introduces a revolutionary change that nevertheless sounds familiar. The phrase provides more friction than illumination because the type of analysis it describes is neither distant, nor is it reading.<sup>14</sup> On the other hand, Michael Gavin’s more recent and more accurate terminology, “literary mathematics,” may send some early modernists scurrying for the exits.<sup>15</sup> It shouldn’t. If we are willing to devote ourselves to understanding an age in which “numbers” could be a synonym for “verse,” and in which every schoolboy learned how to count syllables, it should be no great disruption to our habits of inquiry to embrace the fact that if we want to count more syllables or words than the mind’s ear can hold at any one time, then we are going to use a computer to do it.

The *Lab* offers a range of quantitative modes of access into the *EarlyPrint* data, and while aficionados of statistical methods in linguistic computing will find much that pleases them, the entry points require no advanced mathematical or computational background to get started. For example, armed with categories no more esoteric than “spelling” and “year,” we can discover a great deal about how a particular spelling changed over time using the *N-Gram Browser* available from the *Lab* menu’s “Visualizations” section. The first time we navigate to the browser, it comes pre-loaded with the search

---

<sup>14</sup> Distant reading did indeed reflect a sort of distance in Moretti’s original formulation, one in which he posits, “literary history . . . will become ‘second hand’: a patchwork of other people’s research, *without a single direct textual reading*,” from “Conjectures on World Literature,” *New Left Review* 1 (2000): 67. But by the time of his eponymous book in 2013, Moretti and others were using the term to describe quantitative analysis of primary texts, not the aggregated research of other scholars. Franco Moretti, *Distant Reading* (London: Verso, 2013).

<sup>15</sup> Michael Gavin, *Literary Mathematics: Quantitative Theory for Textual Studies* (Stanford: Stanford University Press, 2023). The phrase “cultural analytics” also has some currency as a description of quantitative methods in the humanities; for a foundational document, see Lev Manovich, “The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics,” *Journal of Cultural Analytics* 1.1 (2016), <https://doi.org/10.22148/16.004>.



## The Spenser Review

terms “love, loue” and “unigrams” with original spellings chosen as input parameters. A unigram is simply an  $n$ -gram where the value of  $n$  is one, i.e., a single word, and by choosing original spellings we are asserting a preference to search word forms that have not been standardized, thus allowing us to see in the plot at the bottom of the screen, where relative frequency is the vertical axis, how the older spelling “loue” is on the rise in usage into the 1520s, levels off for most of the sixteenth century, is supplanted by the modern form “love” in about 1630, and is essentially gone by the early 1640s (see fig. 7).

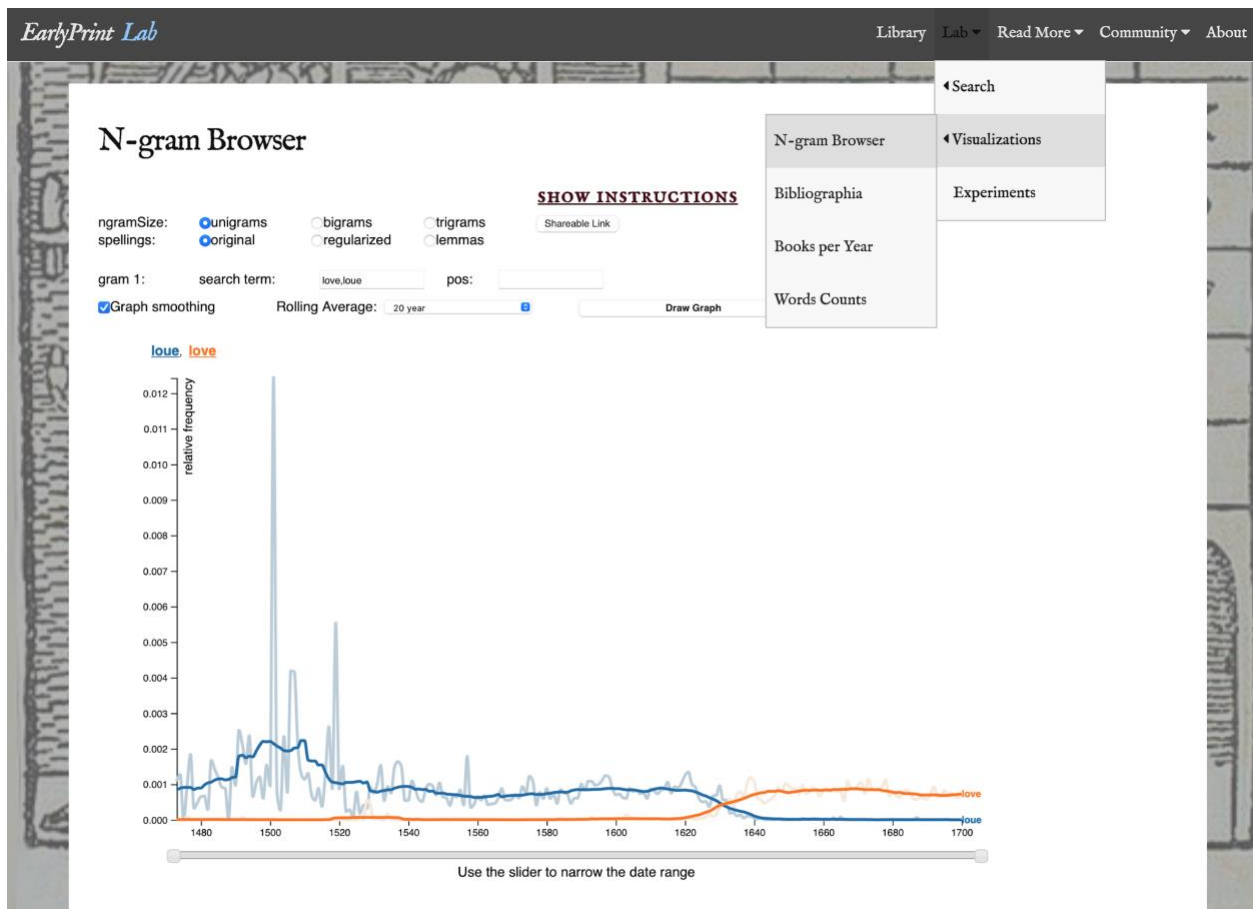


Fig. 7. “love/loue” over time in the “N-Gram Browser.”

This brief example of “loue/love” by no means exhausts what can be done with spelling; Loewenstein and his colleague Anupam Basu have delved considerably deeper into spelling changes over time with their article-length demonstration that Spenser’s spelling eccentricities are not as distinctive as Spenserians have often supposed.<sup>16</sup> And spelling is

<sup>16</sup> Anupam Basu and Joseph Loewenstein, “Spenser’s Spell: Archaism and Historical Stylometrics,” *Spenser Studies* 33 (2019): 63-102. For a more general account of tracing early modern orthographic change and variation, see Anupam Basu, “‘Ill shapen sounds, and false

## The Spenser Review

only one axis of inquiry; the *N*-Gram Browser can perform similar operations on lemmata and parts of speech while handling two- and three-word sequences as well as unigrams.<sup>17</sup>

The *Lab* menu offers four different tools in its “Visualizations” section (of which the *N*-Gram Browser is only one), five different options under the “Search” section, and about ten links in a section called “[Experiments](#).” I will need to limit further discussion to two search options, and merely mention that while the entire *Lab* promotes a spirit of experimentation, the “Experiments” section showcases work that is less finished and less well documented than the rest of the site but suggests possibilities and directions for future tools and analytical endeavors. The most basic search option, “Corpus Search,” provides a sophisticated take on Key Word In Context (KWIC) word searching. Words or word attributes (or sequences of them) may be searched via a graduated set of interfaces that range from very simple with limited control to very powerful at the expense of some complexity. Results may be grouped by words or word attributes that appear immediately before or immediately after the search term, and the list of results not only displays context for each hit but also has a link to the relevant page of the book in the *Library*.<sup>18</sup> C. S. Lewis, no disrespecter of dictionaries, nevertheless noted long ago that, “One understands a word much better if one has met it alive, in its native habitat.”<sup>19</sup> The “Corpus Search” feature of the *EarlyPrint Lab* allows anyone to become an expert word hunter across all billion and a half words of the corpus, easily locating even the rarest quarry and viewing it along with as much or as little of its native habitat as desired. For example, a simple search for “end my song” turns up thirty-eight hits, ten of them, unsurprisingly, in Spenser’s *Prothalamion*, with other occurrences scattered among George Gascoigne, George Wither, John Harington, and some other authors that perhaps even the erudite Lewis would not have recognized. The network of authors and texts in which a word or phrase reappears surely represents a part of its lexical habitat every bit as much as does the immediate context.

The reappearance of a simple phrase may or may not indicate any pervasive similarity among different texts or different authors. For that, we would need to turn to

---

orthography’: A Computational Approach to Early English Orthographic Variation,” in *Digital Turn*, 167-200.

<sup>17</sup> For a fuller explanation, see “Using the N-gram Browser,” [https://earlyprint.org/how-to/howto\\_ngram\\_browser.html](https://earlyprint.org/how-to/howto_ngram_browser.html), and also, “Why N-grams?,” [https://earlyprint.org/how-to/intro-to-ngram\\_browser.html](https://earlyprint.org/how-to/intro-to-ngram_browser.html).

<sup>18</sup> For more details on using this tool, see, “Introduction to Corpus Search,” [https://earlyprint.org/how-to/intro\\_to\\_linguistic\\_search.html](https://earlyprint.org/how-to/intro_to_linguistic_search.html), and “Filtered Search: Early English Poetry,” [https://earlyprint.org/how-to/filtered\\_search.html](https://earlyprint.org/how-to/filtered_search.html).

<sup>19</sup> C. S. Lewis, *Studies in Words*, 2nd ed. (Cambridge: Cambridge University Press, 1967), 2.

## The Spenser Review

“[The Discovery Engine](#)” under the Search section of the *Lab* menu which, given a single text, finds other, similar texts using either relative word frequencies or structural mark-up (such as the division of a play into acts and scenes) as its inputs. The output is a list of texts in which the selected features have been reduced to a similarity coefficient indicating how similar each text is to the source text, ordered from most similar to least. Entering the *Prothalamion*, for example, tells us that, by one word frequency measure, William Drummond has two works that appear in the top ten list of texts most similar to Spenser’s, and by another, Michael Drayton has three (see fig. 8). If we use text structure rather than word frequency, we would see that Spenser himself, via the inclusion of his own *Complaints* and *Colin Clouts Come Home Againe*, is the author whose appearance in the top ten most similar list suggests an affinity with *Prothalamion*. Does this mean that the structure of Spenser’s verse is more distinctive than his vocabulary? It might, but this introductory essay has no intention of spoiling the fun for those who would like to use “The Discovery Engine” to take this line of inquiry further and make their own discoveries.<sup>20</sup>

---

<sup>20</sup> For guidance on use of The Discovery Engine, see “Introduction to the Discovery Engine,” <https://earlyprint.org/how-to/intro-to-disco-engine.html>. For a gentle introduction to the mathematics of similarity, see John R. Ladd, “Understanding and Using Common Similarity Measures for Text Analysis,” *Programming Historian* 9 (2020), <https://doi.org/10.46430/phen0089>.

**The Discovery Engine**  
An experimental tool for discovering similarities between texts

Find the TCP ID for a text      Or, find texts similar to a TCP id

Author:       TCP ID:

Title:       n results:

TF/IDF Results	LDA (Mallet) Results	XML Tag Results
± 1.000 A12781 1596	Spenser, Edmund, 1552?-1599.	Prothalamion or A spousesall verse made by Edm. Spenser. In honour of the double mariage of the two honorable
± 0.957 A67892 1602	Ovid, 43 B.C.-17 or 18 A.D. Beaumont,	Salmacis and Hermaphroditus Salmacida spolia sine sanguine & sudore. Metamorphoses. Salmacis and
± 0.948 A20824 1594	Drayton, Michael, 1563-1631.	Ideas mirrour Amours in quatorzains.
± 0.943 B11895 1595	E. C., Esquire.	Emaricdulfe Sonnets written by E.C. Esquier.
± 0.941 A17042 1625	Browne, William, 1590-ca. 1645.	Britannia's pastorals. The first booke
± 0.940 B00581 1633	L. P. (Laurence Price), fl. 1625-1680?	A compleate gentle-vvoman described by her feature, her person slender, her beauty admirable, her wit excellent,
± 0.939 A20813 1595	Drayton, Michael, 1563-1631.	Endimion and Phoebe Ideas Latmus. Phoebus erit nostra princeps et carminis author.
± 0.937 B05345 1650	S., S.	Loves mistresse or Natures rarity. To a most excellent delicious new tune, called, Tell me you wandring spirits in
± 0.933 A17871 1607	Campion, Thomas, 1567-1620.	The discription of a maske, presented before the Kinges Maiestie at White-Hall, on Twelfth Night last in honour
± 0.932 A20823 1593	Drayton, Michael, 1563-1631.	Idea the shepheards garland Fashioned in nine eglogs. Rowlands sacrifice to the nine Muses.

Fig. 8. Texts similar to Spenser’s *Prothalamion* in “The Discovery Engine.”

What’s next for *EarlyPrint*? In some ways the project is essentially complete and provides unique and unprecedented access to digital surrogates of the surviving record of early English printed materials, but in other ways it is just getting started. At a minimum, there will be more of the same, which means steadily improving transcriptions and thus also linguistic tagging, which will in turn improve search and analysis capabilities. As more rare book libraries serve up scanned books via IIF, more matching page images will be available, and better page images will facilitate more error correction, in a virtuous cycle of steady improvement. But more of the same will be a baseline, not a limit. The digital environment in general and the interests and capabilities of humanities scholars in particular are undergoing noteworthy changes. Natural language processing has come a considerable distance since *MorphAdorner* was developed, and newer techniques, such as Bidirectional Encoder Representations from Transformers (BERT), hold promise for morphological identification, linguistic analysis, and text correction.<sup>21</sup> Current artificial

<sup>21</sup> Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805* (2018), <https://arxiv.org/abs/1810.04805>. For a more accessible introduction, see, "BERT (Language Model)," Version 5, June 30, 2021, *Devopedia*, <https://devopedia.org/bert-language-model>.

## The Spenser Review

intelligence capabilities have progressed to the point that it will soon seem quaint to wonder whether the AI can produce or detect plagiarism, or whether it can write cringeworthy Shakespearean verse. As *EarlyPrint* project team member John Ladd notes in a recent podcast about AI for the Folger Shakespeare Library, “there are many challenges to historical language research and historical language analysis that have posed problems for more traditional, natural language processing. Those problems might be getting a lot easier to handle now. Things like how widely spelling varies across the 17th century, for instance.”<sup>22</sup> Emerging techniques thus offer new ways to answer old questions, where “old” may refer either to the world of pre-digital literary studies that continues and even flourishes alongside later developments, or to the previous iterations of numerical methods and digital representations. In any case, by treating the English texts produced by early printing presses not only as texts but also as data, the *EarlyPrint* project embraces the state of the art of digital mediation as it currently exists, and also lays the groundwork for whatever is around the corner.

---

<sup>22</sup> Barbara Bogaev et al., “Artificial Intelligence Goes to English Class, with Jennifer Black, John Ladd, and Laura Turchi,” *Shakespeare Unlimited* Episode 204 (Washington: Folger Shakespeare Library, February 27, 2023), <https://www.folger.edu/podcasts/shakespeare-unlimited/artificial-intelligence/>.