Digital Studies /
Le champ numérique

Dobson, James, and Scott Sanders. 2022. "Distant Approaches to the Printed Page." *Digital Studies/Le champ numérique* 12(1): 5, pp. 1–28. DOI: https://doi.org/10.16995/dscn.8107

Open Library of Humanities

Distant Approaches to the Printed Page

James Dobson, Dartmouth College, US, james.e.dobson@dartmouth.edu
Scott Sanders, Dartmouth College, US, scott.m.sanders@dartmouth.edu

Laurence Sterne's novel – The Life and Opinions of Tristram Shandy, Gentlemen – includes a tongue-and-cheek moment that prefigures distant reading. Near the end of the sixth volume, the narrator represents uncle Toby's story as a meandering line with unexpected twists and predictable turns. The narrator's precise line is inserted between two paragraphs. Its shape reminds the reader of the book's tangential plot. But it also brings the reader back to the material contours of the story. It is a story that comes into being from the organization of lines and paragraphs on the printed page. The narrator's precise line exists as a material object, in the middle of page 407 in volume 6 of the 1762 Lynch edition. The line gestures towards the physical space that it inhabits. In order to interpret its contours, the reader should also take into account the shape, organization and size of the printed page. This type of material analysis is under-represented in computational humanities, the majority of which has addressed segmented objects at the level of the book—actually, at the level of collections of books. The most common category of this text segmentation procedure is natural to literary scholars: the separation of individual works from within a larger collection of texts. Other categories or types of text segmentation might include the segmentation and parcellation of a longer text into its component chapters or automated algorithmically-defined procedures that ignore chapter and paragraph boundaries to cut a text or collection of texts into equally sized units of words. Segmentation enables comparison of textual objects to determine smaller effects—signals that within the larger stream of words might otherwise be lost. There has been some interest in examining individual sentences. Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, Amir Tevel, and Irena Yamboliev argue that "style" exists at the level or scale of the sentence. Thematic units, however, as Mark Algee-Hewitt, Ryan Heuser, and Franco Moretti argue, might be best captured at the level of the paragraph. Sentences and paragraphs are two different units of segmentation that are both connected with linear, human reading practices. However, segmenting a text into paragraphs rids us of information about the appearance of the paragraph and its relation to the rest of the page remains occluded. Where, for example, does a particular paragraph appear in the space of the page? Are there gaps between paragraphs? Are there printed ornaments, illustrations or annotations? When digital humanists erase the footnotes from Walter Scott's novels, the marginalia from Bunyan's Pilgrim's Progress and the irreverent experimental pages from Tristram Shandy, they lose the page-level context with which these texts are presented.

Le roman Vie et Opinions de Tristram Shandy gentilhomme de Laurence Sterne inclut un momentironique qui préfigure la lecture à distance. Vers la fin du sixième tome, le narrateur décrit l'histoire de l'once Toby comme une ligne sinueuse avec des rebondissements inattendus et des tournures prévisibles. Les lignes précises du narrateur sont insérées entre deux paragraphes. Sa forme rappelle la tangente de l'intrigue au lecteur. Mais aussi, elle rappelle le lecteur du contour matériel de l'histoire. C'est une histoire qui voit le jour à partir d'une organisation de lignes et paragraphes sur des pages imprimées. Les lignes précises du narrateur existent en tant qu'objet matériel, dans le milieu de la page 407 du volume six de l'édition Lynch de 1762. Cette ligne fait signe à l'espace physique que celle-ci habite. Afin d'interpréter ces contours, le lecteur doit alors tenir compte de la forme, de l'organisation et de la grosseur de la page. Ce type de matériel d'analyse est sous-représenté dans le domaine des humanités informatiques, dont la majorité s'adresse à des objets segmentés au niveau du livre – et même au niveau des collections de livres. La catégorie la plus commune de cette procédure segmentée est naturelle pour les spécialistes littéraires : la séparation de travaux individuels au sein de collections de textes plus larges. D'autres catégories ou types de textes segmentés peuvent inclurent la division et le morcellement d'un texte long dans son chapitre ou des procédures algorithmiques définies et automatisées qui ignorent les limites de chapitre ou paragraphe et coupent un texte ou une collection de textes en unités de mots de la même taille. Cette division permet la comparaison d'objets textuels pour déterminer de plus petits effets – des signaux qui seraient autrement perdus dans le flux de mots plus large. Il y a de l'intérêt pour examiner les phrases individuelles. Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, Amir Tevel et Irena Yamboliev soutiennent que le « style » existe au niveau ou à l'échelle de la phrase. Toutefois, comme Mark Algee-Hewitt, Ryan Heuser et Franco Moretti soutiennent, les unités thématiques pourraient être mieux capturées au niveau du paragraphe. Les phrases et paragraphes sont deux unités de segmentation différentes qui sont toutes deux connectées aux pratiques humaines de lecture linéaire. Cependant, diviser un texte en paragraphes nous enlève l'information sur l'apparence du paragraphe et les relations au reste de la page demeurent obstruées. Par exemple, où est-ce qu'un paragraphe particulier apparait sur la page? Est-ce qu'il y a des espaces entre les paragraphes? Est-ce qu'il y a des décorations, illustrations ou annotations imprimées sur la page? Lorsque les humanités numériques effacent les notes de bas de page de romans de Walter Scott, les notes marginales de Le Voyage du pèlerin de Bunyan et les pages expérimentales impertinentes de Tristam Shandy, ils perdent le contexte au niveau de la page dans lequel ces textes sont présentés.

Introduction

From John B. Smith's foundational essay "Computer Criticism" (1978) to Franco Moretti's *Distant Reading* (2013), literary critics using computational methods have identified their work as compatible with or carrying on the work of formalism and structuralism. This framing strategy helped introduce computing to literary criticism by familiarizing the categorization and quantification procedures at the core of these methods. Field framing aside, computational approaches in the humanities are primarily concerned with the internal workings of language or the relational qualities of meaning. While more recently developed word embedding models add a greater degree of contextualization, the dominant natural language processing techniques have treated the text as a flow of (mostly) ahistoric de-materialized signifiers. These methods situate meaning in a text's semantics, word frequency, word placement, or relationships among words. What both word embedding and natural language processing models have in common, then, is a logocentric approach that ignores the material context of the text, which is to say its status as a printed object. Text-based or distant reading and computational methods are only a few of many possible modes of formalist literary criticism, and this formalism, as many have argued over the past decades, is a limited form of critique and an incomplete method of understanding. The printed page, to take just one example, is more than a container of words; the construction and presentation of the page are themselves important sources of meaning that have been ignored or suppressed by text mining methods. Book historians have pointed to the importance of what is called print culture as a way to slide the site of analysis from the page to the book and to the network of laborers involved in the editing, printing, and distribution of text.

In this essay, we use computational methods to extract features of the graphic printed page in order to analyze these features both "distantly" and "closely." In so doing, we address an omission in the current text-focused digital humanities methods. We seek to expand the scope of the formalism made possible by computer-assisted interpretation by using computer vision techniques to examine the appearance and construction of the visual page while at the same time pointing to the limitations of such formalist methods. We have in mind a range of reading strategies that resist the logocentricism found in much contemporary digital humanities work. Lisa Marie Rhody, for example, argues that literary critics using computational methods should consider the "cultural, ethical, and political stakes of their observational position" (Rhody 2017, 263) and feminist interpretive practices, including the ekphrastic tradition that calls into question the knowing glance of distant reading by mediating text and image. Andrew Piper, Chad Wellmon, and Mohamed Cheriet's reframing of historical print objects in terms of the page image rather than textuality introduces a set of terms and procedures from Document Image Analysis (DIA) that enables us to join in the call to "expand the scale of evidence considered when making inferences about the past" (Piper, Wellmon, Cheriet. 2020, 367). Our approach shares with these DIA methods a desire to consider the bibliographic page rather than the extracted text that the dominant text-centric computational methods prioritize. We demonstrate a methodology and a workflow for detecting paratextual objects, a type of printed features that are not representable by optical character recognition and thus frequently absent from the computational analysis of digitized books. We will first discuss eighteenth-century print culture and then shortcomings in existing computational approaches to literary history before introducing our methodology. Finally, using Samuel Richardson's *Clarissa* as a case study, we demonstrate the interpretive possibilities of an expanded digital humanities that is less text-centric and more attentive to the materiality of the book as a set of visual features.

At present, much computational humanities scholarship takes as its object of study decontextualized segments of extracted text. One category of segmented text is natural to literary scholars: the individual work. Yet it is much more common to segment works into smaller pieces, parceling a novel, for example, into its component chapters or using automated and algorithmically defined procedures that ignore chapter and paragraph boundaries to slice a text or collection of texts into equally sized units of words. John B. Smith understands such segments to bear a structural relation to each other in which one segment is always contained by a larger segment. "The text," Smith claims, "can be formally segmented in a step-by-step manner such that each higher segment is defined in terms of units at the next lower level, ranging from the character to the entire work considered as a whole and by extension to the corpus" (Smith 1978, 21). This text segmentation process allows researchers to compare textual objects to determine relatively small effects that might otherwise be lost within the stream of words. Yet it also radically decontextualizes the enclosed words in ways that foreclose many modes of critical analysis while reproducing the cultural fantasy of the text as an autonomous object—a fantasy that produced the idea of the self-contained work of art in the nineteenth century.

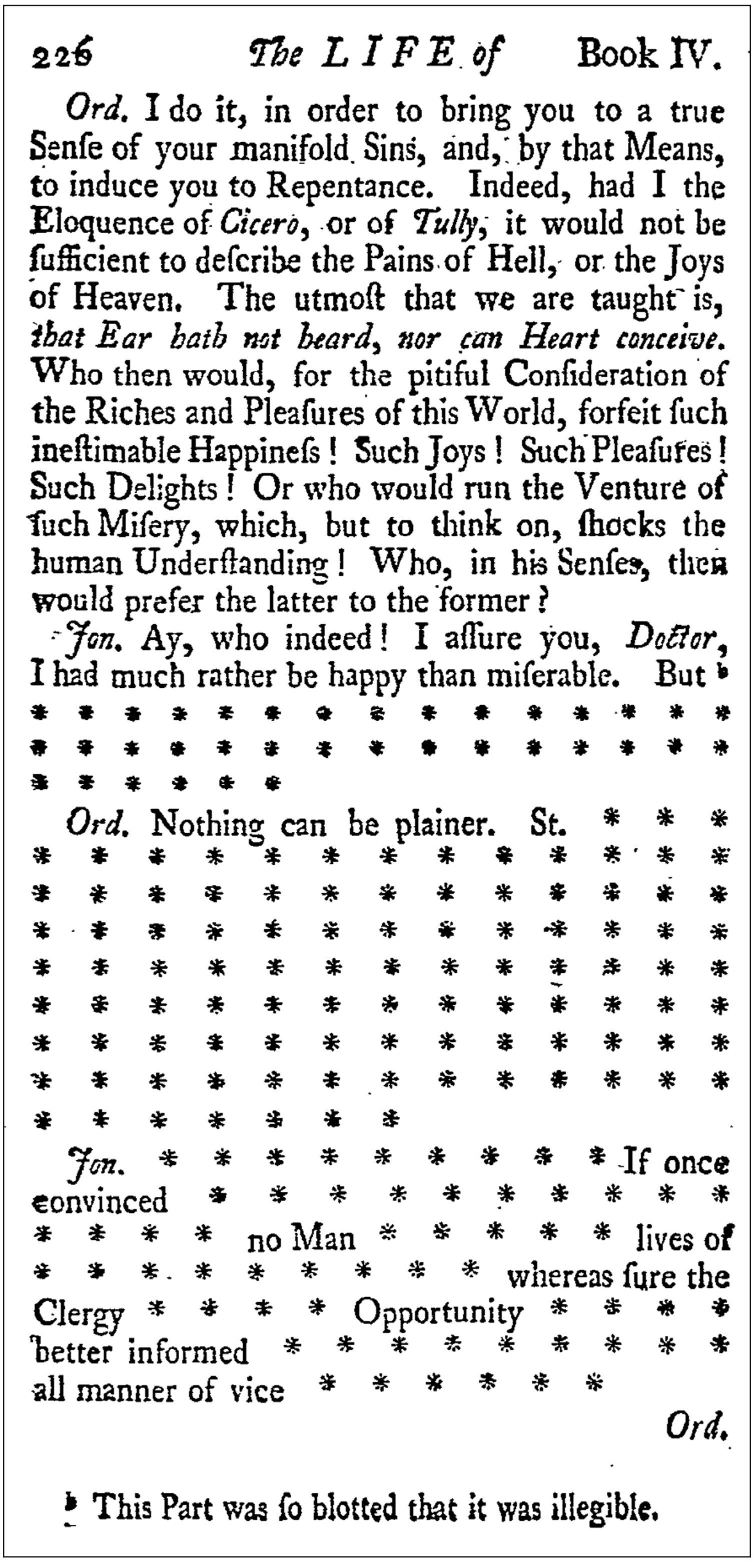
Segmented units of text, especially at the scale of the individual sentence or paragraph, have been used with some success in computational work. Sarah Allison et al. argue that "style" exists at the scale of the sentence, while Mark Algee-Hewitt, Heuser, Moretti claim that "themes" might be best captured at the level of the paragraph (Allison et al. 2017 and Algee-Hewitt, Heuser, Moretti 2017). Because it is connected with linear, human reading, the segmentation of sentences and paragraphs into discrete units seems like a familiar type of formal interpretation. This practice, however, obscures the unit's relation to the rest of the page and the book. How many paragraphs or pages or chapters are squeezed into an arbitrary 1,000-word segment? Where does a particular segmented paragraph appear in the space of the page? Are there gaps between paragraphs? Are there printed ornaments, illustrations, or annotations?

We know that when digital humanists erase the footnotes from Walter Scott's novels, the marginalia from Bunyan's *Pilgrim's Progress*, and the irreverent experimental pages from *Tristram Shandy*, they lose the context in which these extracted and segmented words are presented. What else and who else has been erased?

1. Eighteenth-century print culture

Eighteenth-century authors, as many scholars have argued, were keenly aware of the book as a material object whose appearance invited readers to reflect on its provenance. We have chosen works from this period to illustrate how writers, compositors, and booksellers commented on the creation, meaning, and authenticity of literature through its graphic appearance. In this regard, this period offers an intriguing case study for distant approaches to the page. (The scholarship on eighteenth-century print culture is rich and varied. Our intention, however, is to present a novel computational approach to the digitized page, and thus we do not fully engage with the scholarship on Sterne, Fielding, and Richardson.) Printing involved the transformation of handwritten pages into a printed manuscript. Even three centuries after Gutenberg's invention, writers continued to comment on the transformation of print. For instance, Henry Fielding's satiric novel *The History of the Life of the Late Mr. Jonathan Wild the Great* self-consciously draws the reader's attention to a missing section of dialogue between the Ordinary and Wild (see **Figure 1**). A footnote gestures to a material cause for this lacuna: "This part was so blotted that it was illegible" (Fielding 1743, 226). We highlight this page of Fielding as a moment of self-referentiality in the development of print culture to foreground the attention given to the importance of printing and the visual appearance of paratextual objects during the eighteenth century. Such objects are crucial to understanding the textuality of historical printed texts and regardless of their legibility to popular contemporary text mining techniques, not including these in our analysis produces a gap in our understanding of eighteenth-century visual language.

With digital editions, readers again encounter the question of legibility, which arises from such digital artifacts as poorly scanned page images and incomplete Text Encoding Initiative (TEI) encoded texts. Ryan Cordell uses what he terms "errorful OCR" to foreground the composition of the scanned historical text as a digital text. He asks digital humanists working on digitized printed objects to consider a bibliographic approach that would enable them to "investigate and thus better understand the composition (both technical and social) of the digitized archives they use and to integrate such source criticism into any scholarship that makes claims from the digitized archive" (Cordell 2017, 201). We propose a methodology for interpreting this digital and historical form of re-mediation, first through eighteenth-century approaches to mediation and then through computer vision methods that are capable of understanding the composition of



226 *The* L I F E *of* Book IV.

Ord. I do it, in order to bring you to a true Senſe of your manifold Sins, and, by that Means, to induce you to Repentance. Indeed, had I the Eloquence of *Cicero*, or of *Tully*, it would not be ſufficient to deſcribe the Pains of Hell, or the Joys of Heaven. The utmoſt that we are taught is, *that Ear hath not heard, nor can Heart conceive.* Who then would, for the pitiful Conſideration of the Riches and Pleaſures of this World, forfeit ſuch ineſtimable Happineſs! Such Joys! Such Pleaſures! Such Delights! Or who would run the Venture of ſuch Miſery, which, but to think on, ſhocks the human Underſtanding! Who, in his Senſes, then would prefer the latter to the former?

Jon. Ay, who indeed! I aſſure you, *Doctor*, I had much rather be happy than miſerable. But [b] *

Ord. Nothing can be plainer. St. *

Jon. * * * * * * * * * If once convinced * * * * * * * * * * * * * * no Man * * * * * lives of * * * * * * * * * whereas ſure the Clergy * * * * Opportunity * * * * better informed * * * * * * * * * all manner of vice * * * * * *

Ord.

[b] This Part was ſo blotted that it was illegible.

Figure 1: Henry Fielding, *Miscellanies*, Volume 2, S. Powell, 1743, 226.

the printed page as a visual object. In *This Is Enlightenment*, Clifford Siskin and William Warner argue that the process of mediation is a key method through which to understand the Enlightenment. They present the eighteenth century as "a story of apparent delay in which the fifteenth-century technology of inscription—printing through the use of moveable type—took hundreds of years to implicate and modify an already existing media ecology of voice, sound, image, and manuscript writing" (Siskin and Warner 2010, 10). Eighteenth-century print culture's ambivalent relation to earlier forms of knowledge and cultural transmission, including oral and manuscript forms, were mediated by and registered within printed texts through a variety of medium-specific printed features.

Authors, compositors, and booksellers—those who created and produced printed material —relied on what Lisa Maruca calls the "text work" of print ornaments and paratextual objects, such as the elliptical asterism, to gesture toward their ambivalence about this form of mediation and the losses it threatened. (On "text work," see Maruca 2007, 12–16. While the asterisk ellipsis, as a figure of omission, often refers to textual and oral fragments, this figure has a versatile set of meanings in eighteenth-century texts. In its satirical uses, the ellipsis "represents a form of literary debasement" [Toner 2015, 58]. It can also signify "mental and cognitive failure, argumentative collapse, yawning or sleeping, and a style that can only be associated with the most bawdy parts of the body" [Toner 2015, 58].) J. Paul Hunter, in his reading of *Tristram Shandy*, demonstrates how figures of omission in Sterne's novel (Sterne 1767) comment on print conventions that authors deployed to guide the reading experience (Hunter 1994, 49). Laura Mandell develops this notion further in a discussion of the asterisk, which, she argues, visually encodes its libidinous allusion (Mandell 2007, 763–765). Similar to the asterisks that appear when we enter a password, some of *Tristram Shandy*'s asterisms form a redacted mental image of a word.

The self-consciousness of this period is not merely a reflection on mediation. It also reframes literary production to emphasize the role of the author. As recent scholars have noted, literary self-consciousness has the paradoxical effect of erasing the material context of a book's production. Janine Barchas traces the graphic appearance of the book from the innovation of textual workers to that of certain authors such as Samuel Richardson and Laurence Sterne who attempted to maintain authorial control over their work's *mise-en-page* (Barchas 2003, 11). Lisa Maruca describes a shift from the late seventeenth to the mid-eighteenth century during which booksellers and printers slowly erased the traces of the network of people who produced books. In their place, the author appeared as the proprietor of his literary production, and the good bookseller, such as Robert Dodsley, assisted the author as his literary midwife (Maruca 2007, 10–26). Finally, Christina Lupton explains how "mid-eighteenth-century texts perform through their consciousness of mediation a version of reflexivity that refutes its origins in the human imagination" (Lupton 2012, 11). The literary work thus became a living, breathing organism that was independent of the physical and intellectual labor involved in its production.

Across these recent studies, we see a pattern emerge wherein the people involved in literary production slowly vanish. In their place, the printed book exists as the repository of imaginary characters, as the transcription of the author's ideas, or as an independent work of art. Indeed, Janine Barchas traces this logocentric approach to literature in modern editions, which erase the graphic appearance of the eighteenth-century novel. As she notes, "under the mid-1980s influence of Deconstruction, editorial practice slowly shed its reverence for 'initial' and 'final' intention, a reverence that had traditionally placed an unrelenting emphasis upon first and so-called 'authoritative' editions" (Barchas 2003, 10). In many current computational approaches, we encounter the culmination of this perspective in which a literary work exists outside its material context: islands of words separated from a book's *mise-en-page*, marketing, and consumption. The text alone becomes a worthy signifier of analysis. By devising a distant approach to analyzing the printed page and by examining the page as a constructed visual object, we propose to resituate literary production within the collaborative literary marketplace.

* * *

Some of the more compelling applications of computational methods in the humanities make use of sophisticated machine-learning algorithms to identify potential "signals" within segmented units of texts or to sort texts into well-known, established, or computationally derived categories, such as genres or literary historical periods (Jockers 2013), or contested categories such as the concept of literary prestige (Underwood and Sellers 2016). Data in the digital humanities most frequently means either the text itself or quantitative data derived from statistical measures of the text. Drawing on natural language processing methods developed primarily in computational linguistics and information science, many of these approaches are directed toward such classificatory projects. Given either a set of predetermined formal rules or a set of semantic features that have been algorithmically derived from labeled texts, computational humanists can attempt to classify texts automatically. Poems, for example, can be categorized as elegies, sonnets, or odes by applying additional rules or selecting for specific word, line, or page features.

The distant reading methods used in digital humanities scholarship typically rely on decontextualized text sources in which the text has been stripped of its context within the book—from its appearance on the page to the entire paratextual apparatus, including front matter, title pages, table of contents, colophons, and any back matter. In addition, most digital editions do not clearly indicate or even mention the associated bibliographical metadata, such as publisher, edition, and volume number. The HathiTrust Research Center Extracted Features Dataset, released in mid-2018, marks an important moment in the movement toward standardization of data formats for computational humanists. The present format used by the HTRC dataset offers some page-level features, including term-frequency, line and character counts, and part-of-speech tagging yet not as much attention has been paid to bibliographic metadata. In her critique of the decontextualized and ahistorical "distant readings" offered by prominent digital humanities scholars, including Matthew Jockers and Franco Moretti, Katherine Bode proposes that the field turn to a new object, the digital scholarly edition, to ensure that computational analysis is performed on "carefully historicized texts" (Bode 2017, 93). The vast majority of computational approaches used in the digital humanities treat any printing of a text of interest as essentially equivalent to any other. In the present study, we are not advocating a strictly book history or media studies approach; instead, we want to draw attention to the radical decontextualization of literary texts that has become part of the workflow of most computational humanists.

That said, digital humanities methods are increasing in complexity. Textual sources, in the case of digitized historical sources, first need to be extracted from scanned images of the books in which they are embedded using optical character recognition (OCR) techniques. Following this text identification, extraction, and correction procedure, the input object—the text—generally requires some form of division or segmentation into smaller units for the reduction of input data and for comparison, as we have already mentioned. Franco Moretti refers to both these segments and their sum in his now well-known account of the "little pact with the devil" he calls distant reading: "Distant reading: where distance, let me repeat it, *is a condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: themes, tropes—or genres and systems" (Moretti 2013, 48–49). The digitized text, in Moretti's account, gives way to either smaller or larger segmented units; both of these remediated forms are decontextualizations of the text from the book. But digital humanists do not necessarily need to choose whether to work at the micro or macro scale.

Since 1998, the Text Creation Partnership (TCP) at the University of Michigan has amassed a collection of digital editions, including many from the Early English Books Online (EEBO) collection. While these digital editions are much closer to the scholarly digital edition imagined by Katherine Bode than the plain-text format favoured by many computational humanists, the TCP texts tend to omit bibliographic particularities and ignore many page-level features. Unlike other large-scale collections such as the HATHI Trust archive and Gale's Eighteenth Century Collections Online (ECCO) archive, the texts provided by the TCP are freely available without restrictive research agreements or fees. They also make use of a minimal set of the TEI standard. This encoding makes some forms of automatic text extraction easier—they are certainly preferable to uncorrected OCR for the quality of the text, and the encoding can enable the extraction of dialogue in dramatic texts—but many page-level features cannot be encoded as text.

The Visualizing English Print project (VEP), which makes available data from the XML-based TCP texts in addition to several other archives, illustrates some of the limitations of decontextualized text mining. (A description of the VEP pipeline, including a chart listing all the occurrences of these symbols and marks, can be found on its website and in the code residing in its Github repository [Valenza and Gleicher 2016].) To process, transform, and eventually visualize the data extracted from the texts of interest, the three-stage pipeline presently used by VEP first re-encodes the extended character set of UTF-8 present in the TCP texts into the simplified 128-character ASCII standard. Following this encoding, the VEP pipeline drops all encoding before finally standardizing the spelling of words across its archive. This pipeline is necessary to produce the large number of word occurrences required for some of the higher-level processing of interest to VEP users, including the production of probabilistic topic models. The presence of quantifiable words—relative frequencies or raw word counts of dictionary-corrected words—is all that remains available to interpret the text. Volume, paragraph, and sentence structures remain following this workflow, but page-level structures and any non-ASCII convertible print features are lost. All varieties of hand-encoded asterisms, the print object that is our present concern, are transformed into the asterisk, and many other symbols are completely omitted.

Because many of the available computational methods deal primarily with "raw" text encoded in a simplified plain-text format, many computational humanists think of these segmented units of text as the primary containers of meaning for algorithmic operations (for an account of the "plain text" as format, much like other digital formats, see Tenen 2017). The boundaries of the segments literally "contain" the scope of possible meaning for the included text and/or data. In taking these decontextualized words as the only possible containers of meaning, making them the undifferentiated "bag" in the "bag of words" in which one searches for significance, literary scholars limit their understanding of texts as cultural objects. While Matthew Wilkens, as others have pointed out, suggests that the computational approaches in the humanities should be primarily about extracting information from these containers, he makes possible a scope that exceeds extracted textual features. "We need," Wilkens argues, "to do less close reading and more of anything and everything else that might help us extract information from and about texts as indicators of larger cultural issues. That includes bibliometrics and book historical work, data mining and quantitative text analysis, economic study of the book trade and of other cultural industries, geospatial analysis, and so on" (Wilkens 2012, 251). We agree with Wilkens's call for an expanded scope for cultural analysis, especially his inclusion of book historical work, but we continue to find uses for close reading and remain suspicious of attempts to treat texts and objects as self-evident containers for the extraction of "information" about culture.

Despite Moretti's provocations, scholars who perform automated readings of texts at scale still largely prefer to apply their insights to single volumes. Literary scholars tend to validate models through an explanation of how a particular text fits or does not fit the computational model. (See, for example, Andrew Piper's explanation of the "craft" of computational criticism as directed toward models: "There is understandably an aspect of disenchantment about all of this, as the computational critic lays bare as much as possible of his or her intellectual process. The magic of critical insight is dispelled in favor of the craftsmanship of model-building. But there is something more consensual and less agonistic about the practice of model building as well" [Piper 2018, 11]. Piper shuttles between close reading of individual texts and the results of his model. His close readings work alongside his presentation of data to validate the model for literary scholars.) This preference is the result of disciplinary protocols and training. Literary historians, for example, might favour what Margaret J. M. Ezell calls the "thick description" of the past that preserves and makes available the "details and particularities" found within the archive (Ezell 2017, 16). As literary scholars trained in close reading practices, we know something about the text. Texts are attached to small bits of metadata that are usually well-known: we know the author or authors, we know the year of publication, we even know where the volume was originally published. In segmenting datasets of texts into text-sized units, literary scholars can leverage a whole range of knowledge about the text as an "enclosed" unit that makes the extracted data meaningful within the horizon of meaning enabled by both those bibliographic categories and the "content" of the text itself, including formal knowledge about the narrative and its attendant features.

If automatically segmented textual units, especially standardized, preprocessed chunks of words stripped clean of formal features such as punctuation and line breaks, are the special object of distant reading, then lines, paragraphs, pages, chapters, and books might be considered to belong to the domain of close reading. The return of formalism through computational methods—Moretti's recent use of the term "quantitative formalism" is explicitly designed to present his methods as compatible with traditional literary criticism while at the same time suggesting that his evidence is irrefutable as such—makes up only a small part of the rich set of methods available to critics using close reading for formalist analysis. Any attention to the formal features of objects written expressly for the page (poetry, song lyrics, printed music, figures, and charts, as well as all manner of paratextual objects, including front matter) requires or at least makes available some method of shifting or aligning the extracted textual features, usually words, among multiple digital containers. For example, one might use machine learning algorithms that require segmented lists or "bags" of plain-text words, either in sequential order or not, as input alongside other "expert" rules-based methods that search for prosody and depend upon the presence of all the original words in the correct order and with the original punctuation, spacing, and line breaks.

Close reading and formalist practices at least make it possible to consider the page-level context and the appearance of the words on the page. The now-familiar oppositional framing of close and distant reading is primarily dependent upon an understanding of the link between attention and vision. "Resourceful reading," a concept developed by Katherine Bode to shift the emphasis within quantitative literary studies away from just the results of computation, is more than the combination of close with distant reading. Resourceful reading, Bode and Dixon argue, requires an alignment between the individual text and some of the information extracted from the application of computational methods to datasets and digital archives: "The term 'resourceful reading' was meant deliberately to combine the information-rich, often computational techniques of what has come to be known, after Franco Moretti, as 'distant reading' with close reading's attention to the internal features of individual literary texts: their settings, idioms, themes and patterns of allusion" (Bode and Dixon 2009, 14). We propose that mixed digital methods can be a resourceful method of reading insofar as we can combine attributes of both "close" and "distant" reading and pay attention to omitted dimensions of formalist critique. This does not mean, however, that we are advocating a mode of computational reading that imagines itself as removed from literary hermeneutics and not subject to the norms of critique.

Using machine learning algorithms to classify texts, as we noted above, requires some level of description of the feature selection procedure that can be universally applied to input texts. This means that while we can use extracted sets of word occurrence counts to classify texts—these are, in fact, explicit features, and much more could be said about the process of selecting this particular method and the input texts that make specific feature sets meaningful—the performance of a classifier, as we call this category of machine learning algorithms, will increase if we can provide additional, more meaningful features. In the case of a classification task applied to poetry, some of the more meaningful features contributing to the "correct" classification are not semantic but formal features: indentation, syllable counts, rhyme scheme, length of the lines, stanza length, and count. These features require something a little more complex than the bag-of-words approach applied to automatically extracted segments of words that characterizes much natural language processing work.

The quality of the OCR technology used to transform printed words into digitally encoded representations of words has greatly improved over the last decade. Advances in machine learning technologies combined with improved training data, especially those provided through CAPTCHA problems, have made it possible for algorithms to identify a wide number of printed characters. That said, certain digital collections remain difficult to use in even the most obvious and common kinds of text search and extraction because of lags in updating digital collections with improved OCR technology; random artifacts of the printing process; hand-printed annotations added to the page; the appearance of rare characters, symbols, figures; and difficult-to-render paratextual material. As many scholars have noted, the language and cultural assumptions of many popular OCR techniques are found embedded within the tools. These assumptions, as Alpert-Abrams argues, introduce significant biases about the re-mediated and translated textual signifiers (Alpert-Abrams 2016). Laura Mandell writes of the implications of "dirty" and mistranslated OCR for early modernists: "We are historicists for whom the oddity, the single unexpected usage (the philological equivalent of an anecdote), tells us something of historical relevance, something about how past forms of life harbored different ways of conceptualizing the world than our own" (Mandell 2013, 89). Eighteenth-century texts, especially, present certain complications for OCR procedures, and the extracted text might not be as "clean" as text derived from nineteenth-century and later printing technology. Several scholars have reported mixed results in using ECCO texts for simple queries and keyword searches, never mind the more complex and automated text mining methods (Spedding 2011). While some noise might not prevent text mining operations from extracting useful terms or clusters of words, "dirty" OCR and artifacts introduced in cleaned-up edited digital editions can inadvertently add objects that might become "meaningful" features in training data. (As an example of such spurious correlations, in our initial experiments in using machine learning algorithms to classify poems and songs from the Text Creation Partnership, we discovered that "PDF," a leftover document heading preserved in the TEI/XML files after our extraction of the text, became one of the most statistically significant terms or word features for correctly identifying a ballad as these heading was more frequently found in ballad objects in our training data.)

"Computer vision" is a term of art within computational science that names the automated transformation of image data, including low-level operations such as edge, line, and character detection and high-level procedures such as face detection. Computer vision methods, like the majority of computer methods of interest to humanists, follow the logics of pattern recognition. Samples with some regular distribution of well-defined "patterns" can be used in combination with labeled data—in the case of the previously mentioned applications, samples of typefaces or faces—to detect objects with similar features or patterns. Computer vision methods traditionally operate at the "big data" scale, but unlike text mining methods, computer vision, when applied to images of printed literary texts, can become a form of close reading.

Computer vision methods approach the text as an image rather than encoded text. Scanned pages are projected not in semantic space but in a linear coordinate system of x and y pixel locations and intensity values. Specialized OCR procedures, as a subset of computer vision, can extract text from the page, but these procedures generally ignore both the layout of the text on the page and other non-textual objects, including the paratextual objects invoked above. Because most computational humanists have focused on the extraction of textual information from books, treating digitized texts as images makes it possible to contextualize the text within the book and on the page. Computer vision methods enable a mode of analysis that can combine book historical knowledge, including page-based features, and the extracted text-based methods that dominate computational approaches at present.

In order to correct for some of these limitations in natural language processing techniques, we propose that digital humanists working on historical texts make use of computer vision techniques to locate and extract both known and potentially meaningful features from page images to supplement automatically extracted plain-text sources and expert marked-up encoded texts. Earlier in this essay, we introduce a workflow that uses OpenCV and the Python programming environment to reframe the page as a constructed visual object and thereby facilitate the identification and extraction of paratextual print features. While some research has been done to extract all potential printers' marks in digital archives (Silva 2020), little work has addressed the transformation of these marks and their relation to the printed page. Using samples of well-defined paratextual objects, we searched individual pages for these objects using computer vision pattern-matching techniques. We began our research into the transformation of these objects by analyzing several well-known eighteenth-century texts with what we took to be representative paratextual objects. After identifying these objects, we built a small library of paratextual objects and conducted some simple pattern-matching based searches for these objects. (We have made available our open-sourced code, our paratext image catalog, and a CSV file of all of our extracted page-level features from our selected ECCO image collection in a Github repository: https://github.com/jeddobson/dapp)

2. Case Study: Paratextual Objects in Richardson's *Clarissa*

In Gérard Genette's work on paratext, *Seuils* (Genette 1987), the literary theorist explored, framed and defined the textual objects found within and outside the text. These textual objects include the peritext (titles, chapter titles, table of contents, footnotes) and the epitext (interviews, correspondence). When collated, the peritext and the epitext constitute a work's paratext. For Genette, these textual objects exist at the *seuils* of a text. This French term suggestively associates the paratext with the notion of threshold, entrance, or limit. In this regard, these textual objects function as transitional material that helps the reader enter and exit the text (Genette 1987, 8). A text does not abruptly appear in front of the reader as a collection of sentences, untethered from any framing device. Instead, a text slowly introduces itself to the reader through a series of textual objects that serve as passageways into the text: a reader slowly enters a book through first its title and then its table of contents.

By guiding the reader into the text, paratext also serves as a transactional object, or a sales pitch (Genette 1987, 8). The title invites the reader inside to peruse the narrative's wares. Among the paratextual objects from our study, we found print ornaments, which often preceded a publisher's address to subscribers. These ornaments alerted the reader to upcoming volumes and taught readers to look for portraits and signatures that could help them distinguish the genuine article from a false imprint (Pindar 1795, 43; Skinn 1771, 158). By studying these print ornaments, we begin to see how Genette's initial definition of paratext applies to even these small print artifacts, which create both transitions and transactions between the text and the reader.

Beyond the transactional and transitional functions of paratext, Genette develops five key categories with which to analyze paratext: its placement, historical evolution, form (verbal or nonverbal), addressee, and function (Genette 1987, 10). With these five categories, Genette dedicates hundreds of pages to textual objects, and only at the end of his conclusion does he mention the "immense continent: that of illustration" that he has left unexplored (Genette 1987, 373). Since the publication of *Seuils*, literary scholars have begun to explore non-verbal paratexts that appear in eighteenth-century literature, including illustrations, print ornaments, graphic design elements, musical scores, and annotation. (A number of scholars—Janine Barchas, Nicholas Cronk, Christopher Flint, Ann Lewis, Margaret Linley, Philip Stewart, and Anne Toner, among others—have investigated many non-verbal forms of paratexts.) While recent scholarship has expanded the paratextual universe of Genette beyond the "immense continent" of illustration, scholars' close-reading or historical approaches often focus on a single author's or a series of authors' use of one particular paratextual element. To put an author's use of paratext in the context of the trends of the literary marketplace, however, it is necessary to interpret these objects within the broader context of print culture. To that end, we have experimented with a productive form of "distant" and "close reading" that allows us to contextualize print artifacts.

We developed and refined our paratextual object extraction tool using a partial selection of the literature and language texts found in the ECCO archive. We had access to automatically generated OCR data at the volume level and image file (TIFF) data at the page level for 31,980 volumes. The OCR data is stored in an Extensible Mark-up Language (XML) file and contains paragraph markers and xy coordinates for locating the space occupied by each detected word within the supplied image files. Paratextual objects of the type mentioned above either are not recognized or are incorrectly recognized as words or nonsense ASCII-encoded symbols. The following shows XML mark-up for the first paragraph of the title page of the first volume of the 1719 Taylor printing of Daniel Defoe's *The Farther Adventures of Robinson Crusoe*:

```
<p>
<wd pos=”307,187,443,225”>THE</wd>
<wd pos=”492,183,622,225”>FAR</wd>
<wd pos=”649,181,843,221”>THER</wd>
<wd pos=”124,280,182,362”>A</wd>
<wd pos=”222,273,967,361”>DVENTURE</wd>
<wd pos=”1005,275,1055,360”>S</wd>
<wd pos=”134,450,624,517”>ROBINSON</wd>
<wd pos=”674,448,1022,519”>CRUSOE,</wd>
</p>
```

The images files are variable-sized, compressed TIFF images. By parsing the individual pages and embedded paragraph objects and iterating over individual paragraphs, we were able to extract the "text space" of each page from the XML file. The total page image space was computed by reading the associated TIFF files for each recognized and marked-up page. From these two measurements, we can calculate individual page margins, blank space on the page, and the presence of possible paratextual objects, both known and unknown.

To locate features that are unrepresentable as encoded text from the OCR procedure, we constructed a workflow to extract well-known paratextual objects. We assembled a catalog of these objects and searched for them in the ECCO image dataset by using several popular open-software packages for the Python programming language. We were limited to the instances of the objects in our catalog, with some minor variation in shape and form. For our image processing, we used the Open-Source Computer Vision (OpenCV 3.4.1. 2018) 3.4 package with bindings for Python. The OpenCV package includes a fast tool called the "matchTemplate" function that searches the image space of a target image for the presence of the pixels located in a small sample or patch image.

This search algorithm slides, pixel by pixel, across an image, in our case the image of a page, looking for matches between the template and sections of the target image. Because of the variation found in the type, the amount of ink remaining on the page, the condition of the page prior to imaging, and the artifacts produced as a result of image capture and compression, we needed to produce template images that would "match" with as many appearances of these ornamental marks as possible.

We used as the basis of our search the major paratextual objects found in four editions of Samuel Richardson's *Clarissa*. We expected that the marker called the "inverted asterism," a triangle formed of three asterisks pointing downward, would appear in few other texts, but we found many hundreds of instances of this figure, as well as a number of regular upward-facing asterisms.

We used three major methods to increase the number of matches of our collection of template images. The first included close cropping of images, as seen in the template image of the index or manicule (see **Figure 2**). Smudges at the edges of an image or the use of damaged type that produced an incomplete printed image of an index would still be matched if we provided the smallest number of pixels needed to produce a match in our template image. We also applied several image preprocessing techniques to "smooth" both our template and the target images. This method adds a small amount of distortion or noise to increase the number of pixels that might produce a match. Finally, we experimented with threshold values to lower the number of pixels required to produce a valid match to levels in which we found a very small number of false-positive matches between our template and target images. The eighteenth-century

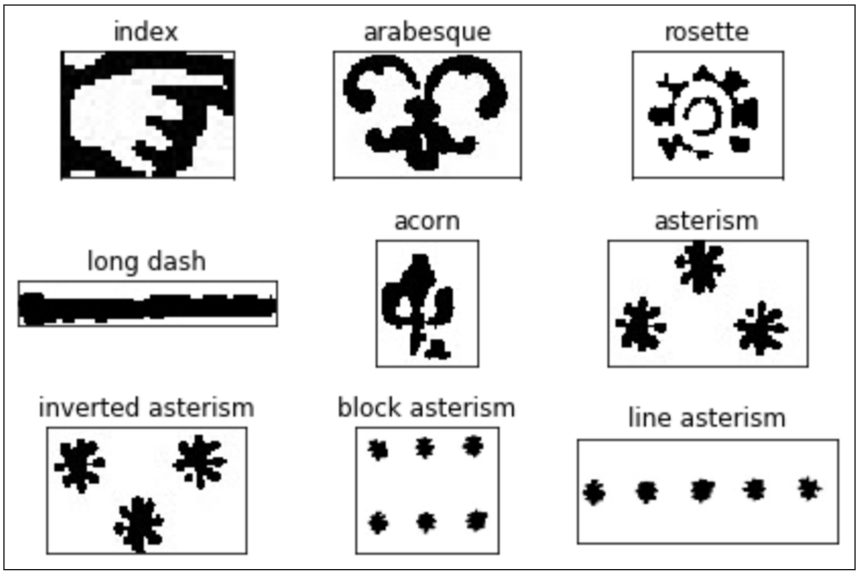
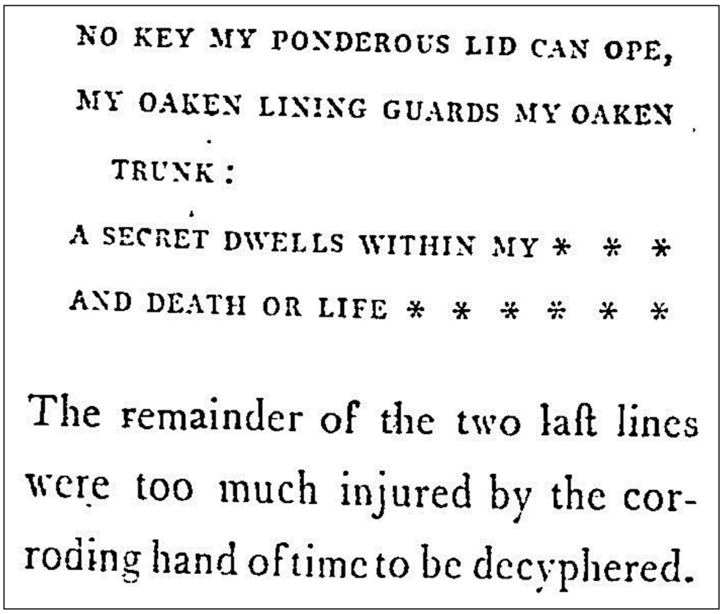


Figure 2: Basic catalog of paratextual ornaments.

books scanned and made available in the ECCO archive contain fair representations of ornamental type, and many of these have been degraded or corroded. **Figure 3** displays an algorithmically detected example of what we called a "line asterism," an eighteenth-century ellipsis. This paratextual marker is used in this instance to represent something unrepresentable from another medium, hand-printed text. Within the fiction of this historical romance, the handwriting has been "too much injured by the corroding hand of time to be deciphered." Our pattern-matching technique was able to detect the compressed and "corroded" print within this remediated representation of an eighteenth-century text stored as a matrix of preprocessed pixel values that had been intentionally deteriorated by our use of image-smoothing algorithms.



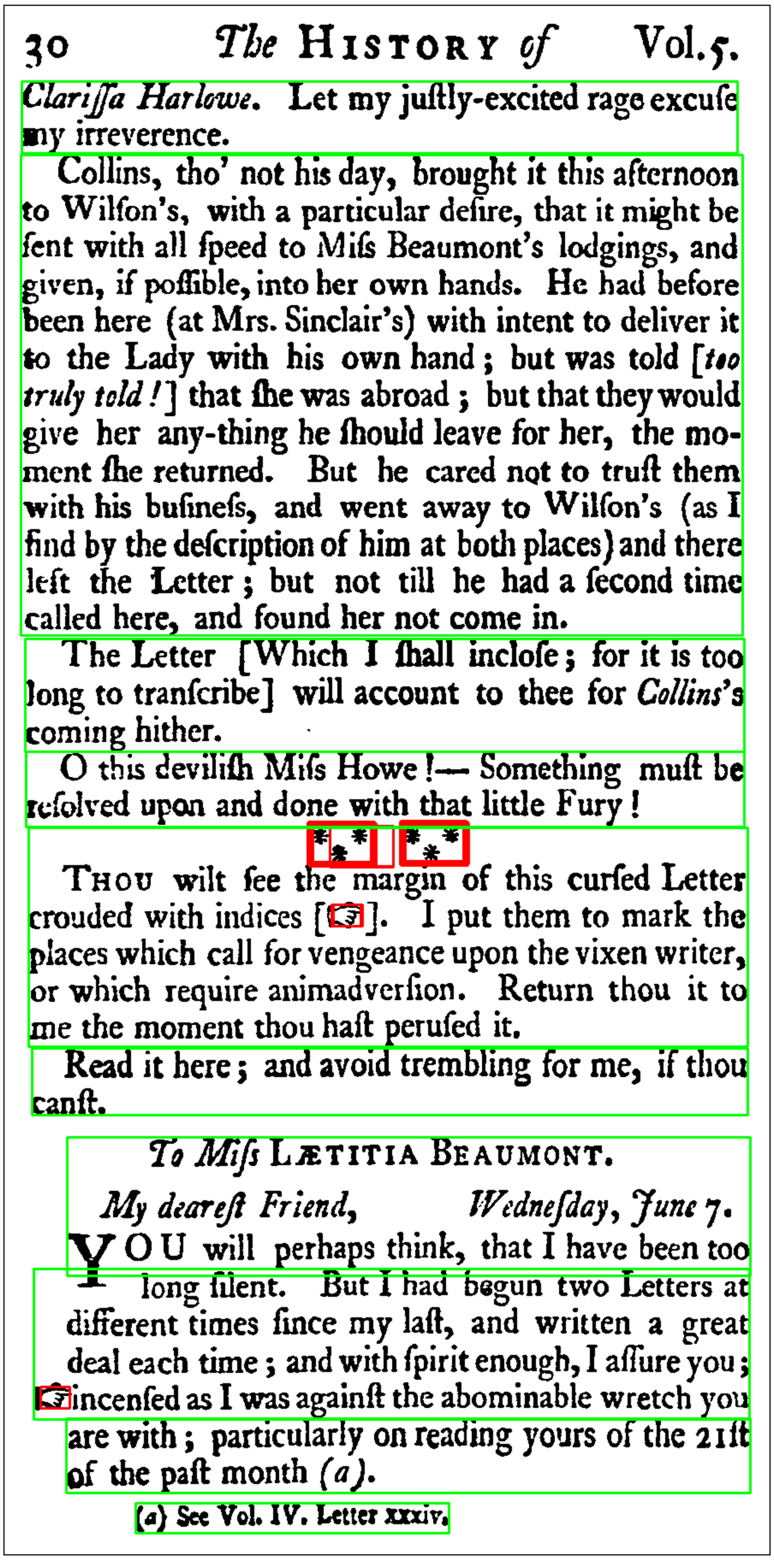
NO KEY MY PONDEROUS LID CAN OPE,
MY OAKEN LINING GUARDS MY OAKEN
TRUNK :
A SECRET DWELLS WITHIN MY * * *
AND DEATH OR LIFE * * * * * *

The remainder of the two laſt lines were too much injured by the corroding hand of time to be decyphered.

Figure 3: John Broster, *Castle of Beeston, or, Randolph, Earl of Chester. An historical romance*, Volume 1, Faulder, 1789, 100.

Computer vision approaches to the digital humanities are both "distant" and "close" forms of computational "reading." Insofar as the approach we have offered makes use of a search for known objects in relation to the page, computer vision might be thought of as a more "formalist" approach than many other computational methods. Since the page is treated as an image, all measurements and features are extracted in relation to other objects on the page. The mixed-method combination of recognized OCR, expert-created XML/TEI marked-up editions, and these automatically extracted features enable us to mark up pages and register or align "corrected" text segments to page images.

In addition to our pattern-matching method, we used other methods to perform additional object detection on book page images. The page image displayed in **Figure 4**



30 *The* HISTORY *of* Vol.5.

Clariſſa Harlowe. Let my juſtly-excited rage excuſe my irreverence.

Collins, tho' not his day, brought it this afternoon to Wilſon's, with a particular deſire, that it might be ſent with all ſpeed to Miſs Beaumont's lodgings, and given, if poſſible, into her own hands. He had before been here (at Mrs. Sinclair's) with intent to deliver it to the Lady with his own hand; but was told [*too truly told!*] that ſhe was abroad; but that they would give her any-thing he ſhould leave for her, the moment ſhe returned. But he cared not to truſt them with his buſineſs, and went away to Wilſon's (as I find by the deſcription of him at both places) and there left the Letter; but not till he had a ſecond time called here, and found her not come in.

The Letter [Which I ſhall incloſe; for it is too long to tranſcribe] will account to thee for *Collins*'s coming hither.

O this deviliſh Miſs Howe!— Something muſt be reſolved upon and done with that little Fury!

* * *   * * *

THOU wilt ſee the margin of this curſed Letter crouded with indices [☞]. I put them to mark the places which call for vengeance upon the vixen writer, or which require animadverſion. Return thou it to me the moment thou haſt peruſed it.

Read it here; and avoid trembling for me, if thou canſt.

To Miſs LÆTITIA BEAUMONT.

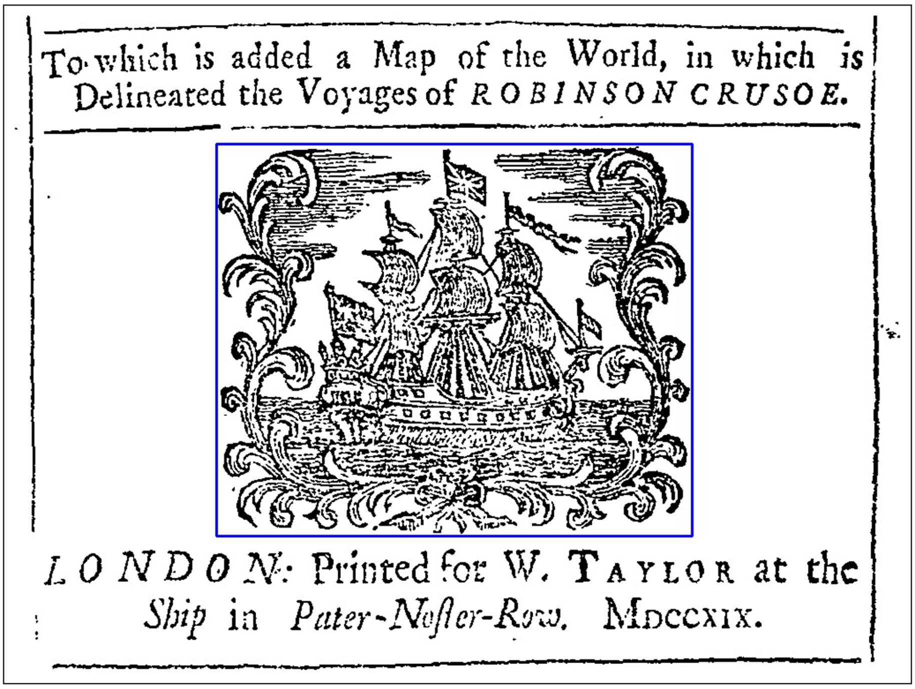
My deareſt Friend, *Wedneſday, June* 7.

YOU will perhaps think, that I have been too long ſilent. But I had begun two Letters at different times ſince my laſt, and written a great deal each time; and with ſpirit enough, I aſſure you; ☞incenſed as I was againſt the abominable wretch you are with; particularly on reading yours of the 21ſt of the paſt month (*a*).

(*a*) See Vol. IV. Letter xxxiv.

Figure 4: Marked-up page: Samuel Richardson, *Clarissa*, Third Edition, Volume 5, 1750, 30.

shows the bounding boxes of OCR text alongside paratextual markers detected in Richardson’s *Clarissa*. We used the pixel location values of the beginning and ending words of each detected paragraph to locate text and then to remove the text for additional image processing. We used this “masking” procedure to remove as much detected text as possible before conducting a search for paratextual objects. Because Richardson places his indices and asterisms within paragraph boundaries, we searched for these before “masking” the detected text. The objects remaining on the page after removing the contents of the bounding boxes—in this case, nothing—were then subject to additional preprocessing and filtering to remove artifacts before they were searched for non-textual objects of interest. We used OpenCV’s edge and contour detection routines to identify such objects. **Figure 5** displays an automatically identified ornamental figure from the title page of Defoe’s *Robin Crusoe*. As in contemporary computer vision applications, such as facial recognition, with enough sample data, ornamental objects can be extracted and identified and machine learning algorithms can be trained to sort through the detected objects, classifying and clustering the results. (For an example of neural-network based computer vision technology being used to match images in the humanities context, see the “Robots Reading Vogue” project at Yale University [King and Leonard 2015].)



To·which is added a Map of the World, in which is
Delineated the Voyages of *ROBINSON CRUSOE.*

LONDON: Printed for W. TAYLOR at the
Ship in *Pater-Noſter-Row.* MDCCXIX.

Figure 5: Bounding box surrounding a detected printer’s mark from the title page to Daniel Defoe, *The Farther Adventures of Robinson Crusoe*, Second Edition, Taylor, 1719, 2.

* * *

Figure 4 displays a few of the printer's ornaments that Richardson inserted into *Clarissa*. As Janine Barchas and Christopher Flint have noted, Richardson first experimented with print ornaments in his first edition, and with the third edition, he associated specific ornaments with characters: the inverted asterism with Lovelace, twin rosettes with Clarissa, and arabesques with Anna Howe (Flint 2011, 131). These print ornaments create a temporal pause in the narrative action and graphically signify the addressee of the text. Janine Barchas, Christopher Flint, and Anne Toner offer slightly different interpretations of these ornaments' temporal function. Janine Barchas interprets them as "signifiers of temporal duration, interruption, and division" (Barchas 2003, 120). Anne Toner argues that they are "like a Hogarthian moral sequence, in that it is between frames—between letters—that narrative action takes place" (Toner 2015, 68). Finally, Flint differentiates each character's ornaments in terms of the narrative action that occurs during the temporal interruption (Flint 2011, 135–37). For Barchas, Toner, and Flint, the ornaments' characterological function extends beyond temporal separation because these ornaments convey symbolic meaning. Elaborating on Richardson's attention to symbols, these scholars extend Terry Castle's and Margaret Ann Doody's analyses of images in Clarissa's coffin text into an interpretation of printed symbols (see Castle 1982, 142–43 and Doody 1974, 186 nl).

From this emblematic perspective, they each introduce compelling nuances into the semiotic significance of Lovelace's inverted asterism. Barchas and Flint symbolically interpret Lovelace's ornament as a constellation of stars, locating the asterisk's meaning in its etymological root ("star"). Barchas's symbolic reading associates Lovelace's starry ornament with his egotism (Barchas 2003, 149), while Flint links it to the "flickering nature" of Lovelace's "volatile relationship" with Clarissa (Flint 2011, 137). Citing J. E. Cirlot's *Dictionary of Symbols*, Flint even suggests that the downward direction of Lovelace's inverted asterism betrays the uplifting symbology of an upward-oriented triangle, and instead represents Lovelace's "downward" and even unholy "impulses" (Flint 2011, 137). Bringing their symbolic reading of the inverted asterism into dialogue with print culture, these scholars contextualize the ornament in relation to the asterism of omission and fragmentation. For instance, Barchas interprets Lovelace's ornament as a symbol of deception because it is reminiscent of inserted elliptical asterisms, which signify textual fragmentation (Barchas 2003, 141). Toner continues the same line of reasoning, framing Lovelace's inverted asterism within her broader analysis of eighteenth-century figures of omission. Citing an eighteenth-century definition of the asterisk, she argues that Lovelace's ornament is a "textual sign" of Lovelace's debased moral character because the asterisk indicates a passage that is "wanting, defective or immodest" (Toner 2015, 76). The framing or reference points selected by the critics, whether they be the literary history of the ellipsis or the etymological origin of the asterisk, necessarily exclude information about the inverted asterism's function in a broader commercial marketplace.

By resituating our frame of reference, "distant reading" enhances traditional literary analysis. For instance, we detected a large number of pages containing inverted asterisms in the ECCO page image collection, and this dataset in turn changed our frame of reference from the symbolic, literary, or ornamental history of the inverted asterism to the bookseller's commercial context (**Figure 6**). In the countless examples we discovered, the inverted asterism appears only occasionally as a separation between sections of text (Addison 1797, 297). In the vast majority of cases, it introduces a publisher's note to the reader. Whether it is found near the beginning or the end of a work, the inverted asterism sits on the threshold between text and reader, publisher and buyer, author and printer. Among the messages that follow the inverted asterism, we find errata, acknowledgments, editorial notes, and annotations (Fitz-Adam 1761, 248, 255). In plays, it alerts the reader to sections that were omitted from a performance (Murphy 1787, 2). As an advertisement, it often previews the upcoming volumes that the bookseller will offer, but occasionally these advertisements tout the medicinal benefits of tonics (Elliot 1770, 1). Finally, on very rare occasions, it even appears as part of a publisher's request for submissions (Anonymous 1785, 223). As an object from the commercial side of publishing, Lovelace's ornament is emblematic of his editorial flair. As Toner notes, "Lovelace is the great linguistic manipulator. He rewrites letters and annotates them with manicules (*or* pointing fingers) to draw out particular messages alien to the writer" (Toner 2015, 70).

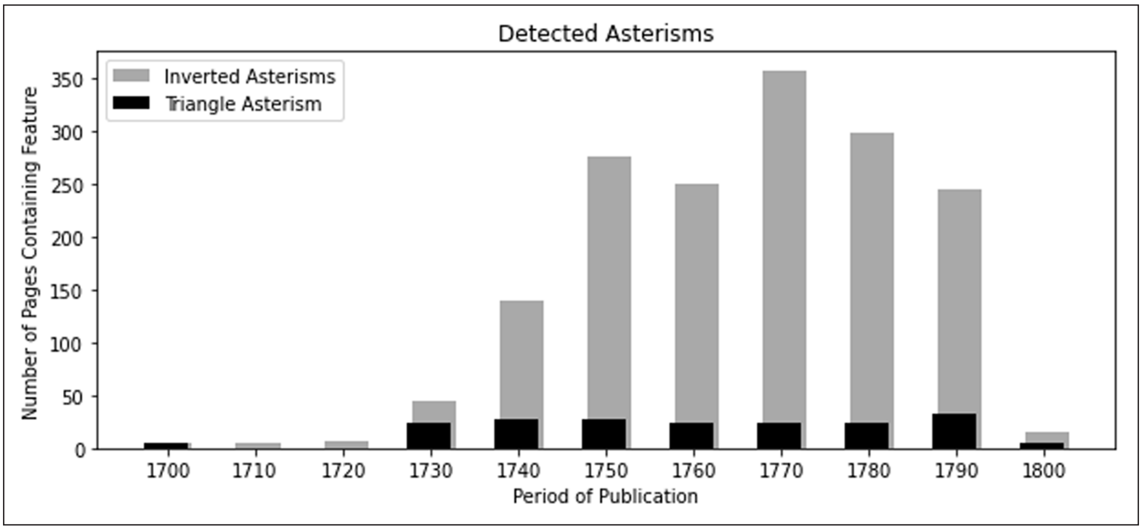


Figure 6: ECCO volume pages with triangle and inverted asterisms from 1700 to 1800.

The inverted asterism and the manicule reveal a more complex typographical character than Lovelace has been thought to possess. As Helen Williams notes, Richardson "experiments with handwriting to heighten the appearance of authenticity" (Williams 2013, 218). "Because Richardson pooled various fonts to raise the hundred or so pieces of type for the margin of the letter, the manicules are not, in fact, identical" (Williams 2013, 214). Going a step further, Marta Kvande explains how Richardson's print techniques imitate manuscript culture not only to convey authenticity, but also to replace manuscript culture with the authority of printed texts (Kvande 2013, 242). As Kvande argues, Richardson's typography "chain[s] self, body, and letter" to such a degree that Clarissa's body is transformed into its graphic representation (Kvande 2013, 242, 245).

By borrowing a symbol that was common within the literary marketplace of the time, Richardson generates a productive tension wherein printed typography imitates hand-written authenticity while also gesturing toward Lovelace's editorial manipulation. In this way, Richardson's printed typography becomes a naturalized emblem of his character. Instead of appearing as marks left by the compositors' hands, these ornaments invite readers to imagine Lovelace's pen drawing his manicules and asterisms, a process that erases, as Maruca notes, "text work" from literary production.

* * *

In interpreting the appearances of the inverted asterism, we are reminded of a poignant commentary from J. Paul Hunter. In his essay on eighteenth-century experiments with print, he encourages scholars to search for innovations outside the literary canon: "But here I want to point explicitly to just one moral for contemporary criticism and theory. In looking for the textual flowerings of technology, historians of texts had better look in more than one place and be ready to see surprising things" (Hunter 1994, 66). Hunter asks historians of material culture, of the book, and of literature to reconsider the traditional frame of reference. We have been arguing that "distant" and "close" reading are already part of humanistic studies. Instead of relying only on computer vision, humanists analyze texts in relation to the interpretive vision of literary critics and theorists. We compare the practices of contemporaneous authors, compositors, and booksellers. We read print manuals and grammar books for insights on paratextual objects. To these acceptable frames of reference, we add human-assisted computer vision.

If the preponderance of algorithmically derived distant readings in the humanities deals primarily with the text as a singular bag of words, the "raw" input device of the plain-text dataset, close readings are associated with the way words appear on the page. Methods that enable the alignment of the encoded text, a page-level image of the historical text, and the plain text are required to address the complexities of historical texts and all the objects embedded and bound within these texts. Computational methods can direct closer attention to the presently unrepresentable formal features of the text. These include measurements of pages and the words on the page, paratextual objects, and the accompanying bibliographic metadata that can help make sense of all these objects.

Turning back to the page, linking page-level representation to positions within the bag of words, might be the best way to give greater attention to the embedded objects within literature. This is not to say that we want to put large-scale text mining aside. On the contrary, our computational methods almost always begin with and depend upon the "distant" algorithmic manipulations of encoded text. But we believe that it is necessary to turn to field expert–produced editions of texts and to additional information gleaned from the pages of our objects of interest. In her theorization of the TEI, Susan Schreibman calls for the emergence of the "encoder-assembler," a figure who produces encoded texts as the equivalent of electronic monographs (Schreibman 2002). Encoded and assembled editions with identified and labeled objects supplement the raw text of the text miner. The best training data for the application of any machine-learning method comes from expert-labeled or expert-tagged data.

We regard the page and the objects located on and designed for the page as some of the most crucially important features for the application of machine learning to historical literary texts. Computer vision-based approaches to the digital humanities allow us to see literary texts with new eyes. The defamiliarizing gaze produced by the algorithmic breakdown of the page image into smaller components can become an important source for feature selection. When combined with text mining, these computer vision approaches might not make the unrepresentable visible, but in the hands of literary critics and historians, they can illuminate previously occluded patterns within and outside the text. The "units" of analysis in computer-aided work in the humanities need to be flexible and porous. The wish for stability seldom matches the reality of our objects: paratexts slide into the text; conceptions of authorship and printing reverse or combine; revisions alter meaning in unanticipated ways; and the same words printed in a different form are fundamentally a different text.

Competing interests

The authors have no competing interests to declare.

Contributors

Authorial contributors

Authors are listed in alphabetical order. Author contributions, described using the *CASRAI CredIT typology*, are as follows:

Author names and initials:
James E. Dobson, Dartmouth College: jed https://orcid.org/0000-0001-8357-7240
Scott M. Sanders, Dartmouth College: sms https://orcid.org/0000-0002-6382-3456

The corresponding author is jed
Conceptualization: jed, sms
Investigation: jed, sms
Methodology: jed, sms
Software: jed
Visualization: jed
Writing – original draft: jed, sms
Writing – review & editing: jed, sms

Editorial contributors

Section Editor:
Iftekahr Khalid, Journal Incubator, University of Lethbridge, Canada

Bibliographies Editor:
Shahina Parvin, Journal Incubator, Brandon University, Canada

Text and Citations Editor:
Morgan Pearce, Journal Incubator, University of Lethbridge, Canada

References

Addison, Joseph. 1797. *The Spectator.* London: H. Baldwin.

Algee-Hewitt, Mark, Ryan Heuser, and Franco Moretti. 2017. "On Paragraphs: Scale, Themes, and Narrative Form." In *Canon/Archive: Studies in Quantitative Formalism from the Stanford Literary Lab*, edited by Franco Moretti, 65–94. New York: N+1.

Allison, Sarah, Marissa Gemma, Ryan Heuser, Franco Moretti, Amir Tevel, and Irena Yamboliev. 2017. "Style at the Scale of the Sentence." In *Canon/Archive: Studies in Quantitative Formalism from the Stanford Literary Lab*, edited by Franco Moretti, 33–63. New York: N+1.

Alpert-Abrams, Hannah. 2016. "Machine Reading the *Primeros Libros*." *Digital Humanities Quarterly* 10(4). Accessed August 10, 2021. http://www.digitalhumanities.org/dhq/vol/10/4/000268/000268.html

Anonymous. 1785. *The Quaker: A Novel, in a Series of Letters*. London: William Lane.

Barchas, Janine. 2003. *Graphic Design, Print Culture, and the Eighteenth-Century Novel*. New York: Cambridge University Press.

Bode, Katherine. 2017. "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History." *Modern Language Quarterly* 78(1): 77–106. DOI: https://doi.org/10.1215/00267929-3699787

Bode, Katherine, and Robert Dixon. 2009. "Resourceful Reading: A New Empiricism in the Digital Age?" In *Resourceful Reading: The New Empiricism, eResearch and Australian Literary Culture*, edited by Katherine Bode and Robert Dixon, 1–27. Sydney: Sydney University Press.

Castle, Terry. 1982. *Clarissa's Ciphers: Meaning and Disruption in Richardson's "Clarissa."* Ithaca and London: Cornell University Press.

Cordell, Ryan. 2017. "'Q i-jtb the Raven': Taking Dirty OCR Seriously." *Book History* 20: 188–225. Accessed August 10, 2021. https://ryancordell.org/research/qijtb-the-raven/. DOI: https://doi.org/10.1353/bh.2017.0006

Doody, Margaret Anne. 1974. *A Natural Passion: A Study of the Novels of Samuel Richardson*. Oxford: Clarendon Press.

Elliot, Adam. 1770. *A True Narrative of the Life of Mr. George Elliot, Who Was Taken and Sold for a Slave; with His Travels, Captivity, and Miraculous Escape from Salle in the Kingdom of Fez*. London: T. Bailey.

Ezell, Margaret J. M. 2017. "Big Books, Big Data, and Reading Literary Histories." *Eighteenth-Century Life* 41(3): 3–19. DOI: https://doi.org/10.1215/00982601-4130753

Fielding, Henry. 1743. *Miscellanies: Volume 2*. UK: Bristow & Garland. DOI: https://doi.org/10.1093/oseo/instance.00058169

Fitz-Adam, Adam. 1761. *The World*. The Third edition. Volume 2 of 4. London: R. & J. Dodsley.

Flint, Christopher. 2011. *The Appearance of Print in Eighteenth-Century Fiction*. New York: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511996344

Genette, Gérard. 1987. *Seuils*. Paris: Editions du Seuil.

Hunter, J. Paul. 1994. "From Typology to Type: Agents of Change in Eighteenth-Century English Texts." In *Cultural Artifacts and the Production of Meaning: The Page, the Image, and the Body*, edited by Margaret J. M. Ezell and Katherine O'Brien O'Keeffe, 41–69. Ann Arbor: University of Michigan Press.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana, Chicago, and Springfield: University of Illinois Press. DOI: https://doi.org/10.5406/illinois/9780252037528.001.0001

King, Lindsay, and Peter Leonard. 2015. "Robots Reading Vogue." Yale University Library. Accessed August 11, 2021. https://dhlab.yale.edu/projects/vogue

Kvande, Marta. 2013. "Printed in a Book: Negotiating Print and Manuscript Cultures in *Fantomina* and *Clarissa*." *Eighteenth-Century Studies* 46(2): 239–257. DOI: https://doi.org/10.1353/ecs.2013.0008

Lupton, Christina. 2012. *Knowing Books: The Consciousness of Mediation in Eighteenth-Century Britain*. Philadelphia: University of Pennsylvania Press. DOI: https://doi.org/10.9783/9780812205213

Mandell, Laura. 2007. "What Is the Matter? Or, What Literary Theory Neither Hears nor Sees." *New Literary History* 38(4): 755–776. DOI: https://doi.org/10.1353/nlh.2008.0008

---. 2013. "Digitizing the Archive: The Necessity of an 'Early Modern' Period." *Journal for Early Modern Cultural Studies* 13(2): 83–92. DOI: https://doi.org/10.1353/jem.2013.0019

Maruca, Lisa M. 2007. *The Work of Print: Authorship and the English Text Trades, 1660–1760*. Seattle: University of Washington Press.

Moretti, Franco. 2013. *Distant Reading.* New York: Verso.

Murphy, Arthur. 1787. *The Grecian Daughter*. London: Lowndes and Bladon.

OpenCV 3.4.1. 2018. Accessed August 11, 2021. https://opencv.org/opencv-3-4-1.html

Pindar, Peter. 1795. *The Lousiad, an Heroi-Comic Poem*. London: George Goulding, James-Street Covent-Garden, and John Walker.

Piper, Andrew. 2018. *Enumerations: Data and Literary Study.* Chicago: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226568898.001.0001

Piper, Andrew, Chad Wellmon, and Mohamed Cheriet. 2020. "The Page Image: Towards a Visual History of Digital Documents." *Book History* 23: 365–397. DOI: https://doi.org/10.1353/bh.2020.0010

Rhody, Lisa Marie. 2017. "Beyond Darwinian Distance: Situating Distant Reading in a Feminist *Ut Pictura Poesis* Tradition." *PMLA* 132(3): 659–667. DOI: https://doi.org/10.1632/pmla.2017.132.3.659

Schreibman, Susan. 2002. "Computer-Mediated Texts and Textuality: Theory and Practice." *Computers and the Humanities* 36(3): 283–293. DOI: https://doi.org/10.1023/A:1016178200469

Silva, Andie. 2020. "Wilkinson, Hazel, principal investigator. Fleuron: A Database of Eighteenth-Century Printers' Ornaments." *Renaissance and Reformation/Renaissance et Réforme* 43 (1). Principal Investigator – Hazel. Accessed August 11, 2021. DOI: https://doi.org/10.33137/rr.v43i1.34091

Siskin, Clifford, and William Warner eds. 2010. *This Is Enlightenment*. Chicago: University of Chicago Press.

Skinn, Ann Emelinda. 1771. *The Old Maid; or, History of Miss Ravensworth*. Volume 3. London: J. Bell.

Smith, John B. 1978. "Computer Criticism." *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, edited by Rosanne G. Potter. Philadelphia: University of Pennsylvania Press, 1989, 13–44. Originally published as "Computer Criticism." *Style 12* (1978): 326–356.

Spedding, Patrick. 2011. "'The New Machine': Discovering the Limits of ECCO." *Eighteenth-Century Studies* 44(4): 437–453. DOI: https://doi.org/10.1353/ecs.2011.0030

Sterne, Laurence. 1767. *The Life and Opinions of Tristram Shandy, Gentleman- Volume 6*. London: Lynch. DOI: https://doi.org/10.1093/oseo/instance.00167758

Tenen, Dennis Yi. 2017. *Plain Text: The Poetics of Computation.* Stanford: Stanford University Press. DOI: https://doi.org/10.1515/9781503602342

Toner, Anne. 2015. *Ellipsis in English Literature: Signs of Omission*. Cambridge, United Kingdom: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139680974

Underwood, Ted, and Jordan Sellers. 2016. "The Longue Durée of Literary Prestige." *Modern Language Quarterly* 77(3): 321–344. DOI: https://doi.org/10.1215/00267929-3570634

Valenza, Robin, and Michael Gleicher. 2016. "Visualizing English Print." Accessed August 11, 2021. https://graphics.cs.wisc.edu/WP/vep/

Wilkens, Matthew. 2012. "Canons, Close Reading, and the Evolution of Method." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 249–258. Minneapolis: University of Minnesota Press. DOI: https://doi.org/10.5749/minnesota/9780816677948.003.0026

Williams, Helen. 2013. "Sterne's Manicules: Hands, Handwriting and Authorial Property in *Tristram Shandy*." *Journal for Eighteenth-Century Studies* 36(2): 209–223. DOI: https://doi.org/10.1111/j.1754-0208.2012.00512.x