Jonathan Reeve

Dennis Tenen

Form, Formula, Format

23 December 2016

A Quantitative Analysis of Critical Quotations from George Eliot's *Middlemarch*

"If we had a keen vision and feeling of all ordinary human life," the narrator of George Eliot's *Middlemarch* argues, "it would be like hearing the grass grow and the squirrel's heart beat, and we should die of that roar which lies on the other side of silence" (Eliot 2004, 180). This sentence is among the most quoted from the novel, appearing in over fifty-six critical articles. What is it about this passage that has made it so quotable? The following describes an experiment in the quantitative analysis of literary criticism surrounding *Middlemarch*, where we attempt to answer this question. Using text reuse detection techniques, we identify the passages of the novel that are most cited by critics, and patterns in the way these passages shift over time. We study the differences in citations between George Eliot specialists and other literary critics, along with other such demographic trends. Finally, we look for patterns in the language of the quotations themselves, in an attempt to discover syntactic and lexical criteria that make a passage likely to be cited.

This study is by no means the first of its kind. Goldstone and Underwood (2014) analyze 21,367 articles spanning over a century and seven literary critical journals. Among their observations are a decrease in the use of numbers over the twentieth century, and an increase in themes of death and violence. The Viral Texts project (Smith, Cordell, and Dillon 2013) maps journalistic text reuse in 19th-century newspapers, and the recent Digital Breadcrumbs of Brothers Grimm project computationally detects text reuse of Brothers

Grimm material. Dennis Tenen (2017) has recently used extracted citations in an analysis of the journal *Comparative Literature*. JSTOR Labs, in their Understanding Shakespeare and Understanding the Constitution projects, construct electronic editions of these texts that denote the number of quotes, and the quoting entities, from each journal in their database. We are indebted to each of these projects for their methodological ingenuity.

George Eliot's novel *Middlemarch* is an unusually apt novel for this study, since as a canonical text that is so self-consciously aphoristic, it is both highly quotable and often quoted. Leah Price notes that it has been "more ruthlessly excerpted than any [novel] since: chopped into anthology-pieces, recycled as calendar decorations, used to test army officers, deployed in a Zionist tract, plastered onto billboard, and quarried for epigraphs to a socialist treatise and even an abridgement of Boswell's *Life of Johnson*" (2003, 9–10). It has been the star, or victim, of at least three nineteenth-century collections of Eliot's aphorisms: *Wise, Witty, and Tender Sayings in Prose and Verse Selected from the Works of George Eliot* (Eliot 1873), the *George Eliot Birthday Book* (Eliot 1878), and *A Moment Each Day with George Eliot* (Moore 1903), the first two of which are described in detail in Price. The editor of *Wise, Witty, and Tender Sayings*, Alexander Main, declares in his preface that Eliot's work "can never again be regarded as mere story-telling … she has for ever sanctified the Novel by making it the vehicle of the grandest and most uncompromising moral truth" (Eliot 1873 x). It is this "sanctified" quality of Eliot's excerpts that we will try to understand with the following analysis.

1. The Corpus

Through the generous help of JSTOR Labs, we were able to obtain a corpus of all 6,069 articles in the JSTOR database that use the word *Middlemarch*. Typically such keyword searches would be too imprecise, but it worked to our advantage that the fictional village name *Middlemarch* rarely, if ever, appears in contexts other than those discussing the novel. These articles are written in twelve languages, including Irish and Welsh, were published in 365 different journals, and represent the work of 4,231 critics[1]. The chronological distribution of this corpus is heavily skewed right: there are 1,400 of these articles from the 2000s decade, but fewer than 100 from the 1940s. Some of the diachronic analyses to follow, therefore, should be treated with a certain amount of skepticism, since the available data are not uniformly distributed. The journals most represented here are, in order of frequency, *Victorian Studies*, *George Eliot - George Henry Lewes Studies*, *Nineteenth-Century Fiction*, *The Modern Language Review*, and *The Review of English Studies*.

In addition to the articles themselves, JSTOR was kind enough to share the results of their topic modeling experiments with these articles. Topic modeling—a statistical technique that uses probabilistic models to infer topics, or clusters of co-occurring words, from documents—has often been used for diachronic analyses such as these. David Blei's foundational work, for instance, studies changing topics in over a century's issues of the journal *Science* (2012, 81); Goldstone and Underwood's aforementioned study uses a similar technique. JSTOR Labs not only inferred topics from each of their articles, but labeled each topic using a manually-assembled topic thesaurus. The most frequently occurring topic la-

[1]Since we have no reliable method for deduplicating names with different morphologies, this count, as well as that of the journals, should be regarded as approximate.

bels in our corpus of critical documents are, in this order: *novelists*, *novels*, *pity*, *sympathy*, *irony*, and unsurprisingly, *literary criticism*.

A diachronic analysis of these topics hints at the protean landscape of 20th century literary criticism, as seen through the study of *Middlemarch*. Figure 1 shows seven abstract topic labels selected from the top 30 topics, and the frequency of their occurrence by decade. The most frequently-occurring topic, *pity*, seems to decline steadily from the 1910s until today. *Satire*, as well, seems to reach its apex in the 1930s, declining rapidly thereafter. *Irony*, in contrast, appears in the 1940s, and by the 1970s overtakes *pity* as the most prominent topic. *Love* remains of lesser prominence throughout, but with a conspicuous absence in the early 20th century that may or may not be attributable to missing data. *Beauty* is the most frequently occurring topic of 1920s criticism, but never appears again. The topic label *women* rises in the 1980s, coincident with the rising popularity of feminist criticism, and reappears in every subsequent decade. Of course, these topic labels should not be understood as proxies for human-identified subjects or themes. Neither should they even be understood as reliable computational approximations, since the irregular shape of the corpus undoubtedly skews this analysis. Still, these results are enticing, and warrant further investigation.

Figure 2 shows a more metacritical selection of topic labels from among the top 30. The dominating topic is *literary criticism*, which might speak to a literary critical self-consciousness that begins in the 1920s and peaks in the 1960s, with a marked drop in the 2000s[2]. This is almost exactly the same arc traveled by *bibliographies*, also a potentially reflexive topic. Both *biography* and *anthologies* are multi-modal, seeing sudden peaks and

[2]This trend is very similar to that which Goldstone and Underwood identify in their corpus for the topic they label *criticism*—see Goldstone and Underwood (2014) 371, Figure 4. Since the peak of their topic is in the 1950s, one might hypothesize that *Middlemarch* criticism is more conservative, in the etymological sense, than average.

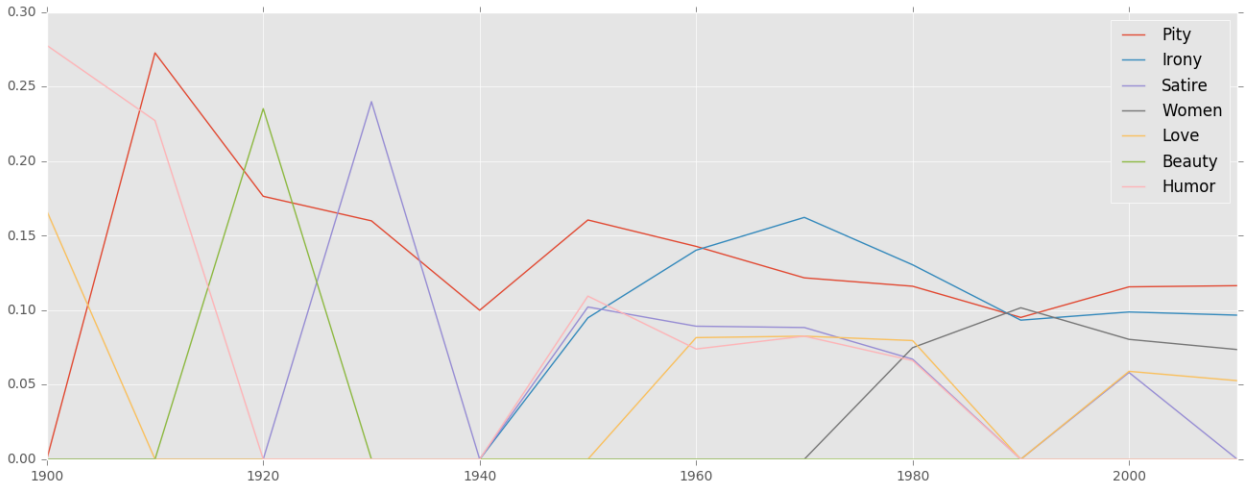

Figure 1: Topics in *Middlemarch* Criticism A

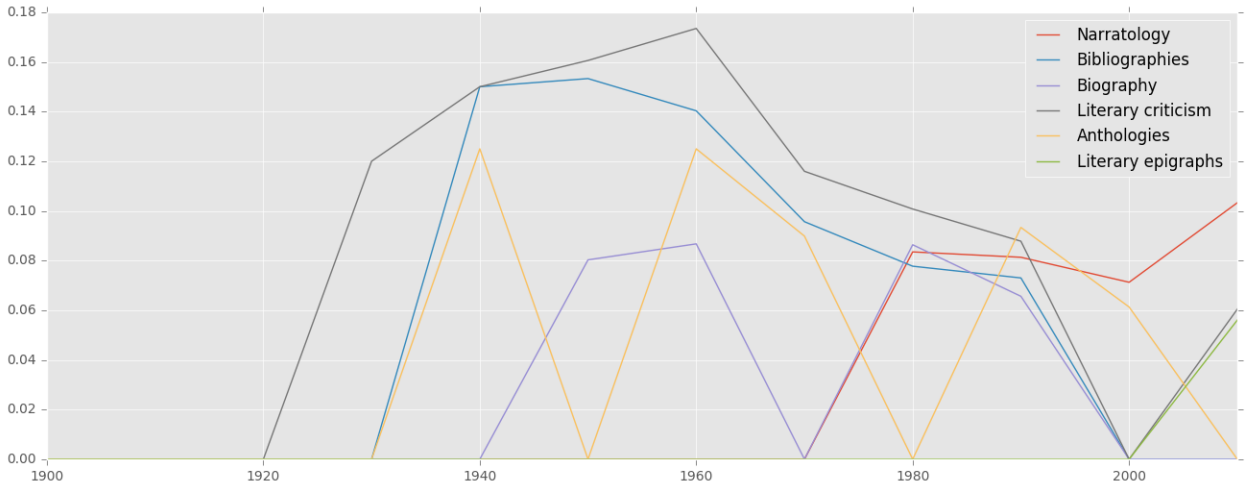

Figure 2: Topics in *Middlemarch* Criticism, B

valleys in regular intervals, perhaps coincident with the fashion of these critical interests. Narratology, a recent brand of literary criticism, arrives in the 1980s, closely following with the Anglophone popularization of Russian formalism. Finally, the *literary epigraphs* topic seems to appear only in the past decade.

It should be noted that the corpus, while almost entirely composed of *Middlemarch* criticism, does contain some noise. One of the more obscure topic labels, with six associated articles, is, inexplicably, *underwater photography*. The corpus also includes at least five

hundred articles with titles like "Front Matter," "Back Matter" and "Index." Rather than prune these articles from the corpus, however, we counted on our text matching algorithm to simply ignore articles without quotations from *Middlemarch*.

2. Analysis

To identify critical quotations from *Middlemarch*, we built a text-reuse detection program, text-matcher, to identify similar passages between *Middlemarch* and critical texts. The program begins by cleaning and stemming words with the Lancaster stemmer, reexpresses them as sets of overlapping trigrams, and then uses the Python difflib to find the longest intersections of those sets. The method used here was adapted from those described in recent text-reuse papers, such as Bär, Zesch, and Gurevych (2012). The program outputs the number of matching sequences per article, along with the character offsets of each match in each document.

Using the relative frequencies of citations per letter, we were able to create a citation-heatmap annotated edition of *Middlemarch*, where each passage is colored according to the number of times it has been cited. Black indicates text that has not been identified by our text matcher in any critical article, and yellow indicates text that has been cited the most often—upwards of two hundred times. This edition provides a useful way of reading the novel, since it literally highlights passages that other readers might highlight. Readers might choose to read only the critically-highlighted passages, for instance, effectively abridging the novel, or they might choose to read only the unhighlighted phrases, to look for areas of the novel that may have been, by this citation metric, critically neglected.

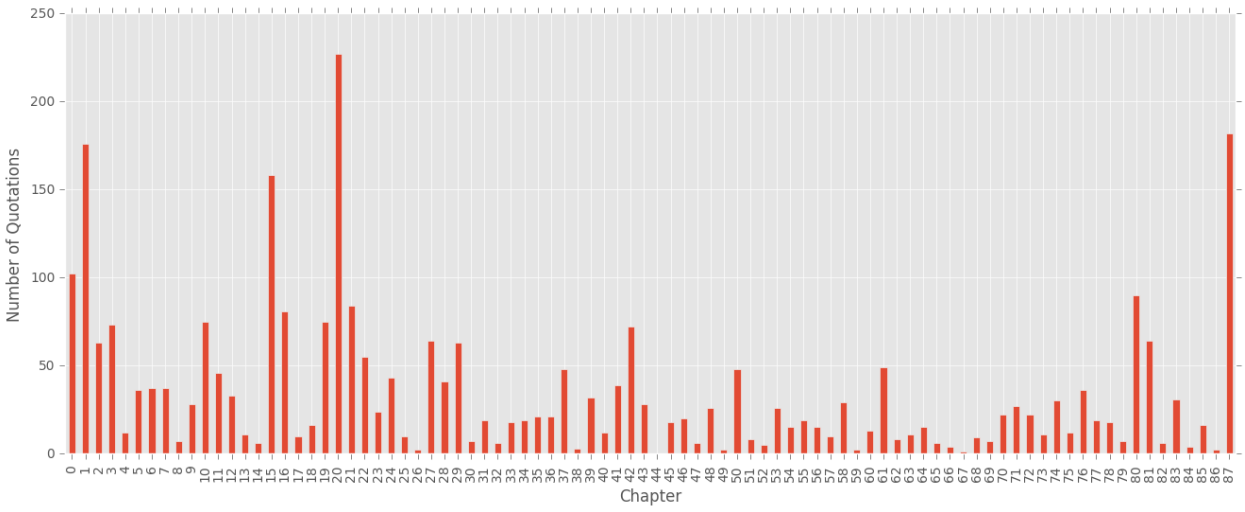

Figure 3: Middlemarch Citations, by Chapter

Figure 3 shows the number of critical quotations identified by the text matcher, quantified by the chapter from which they are quoted. Chapter zero refers to the Prelude, and Chapter 87, since there are only 86 chapters in the novel, refers to the Finale. Some broad trends here are worth noting. Most obviously, the first half of the novel is much more heavily quoted from than the second. When each quarter is summed, the first quarter is shown to have been cited at least twice as much as the second quarter, which is cited twice as much as the third quarter. The fourth quarter sees only a slight increase in citations from the third. The overall pattern is one in which critical interest in this long novel, if it could be expressed in part by this index, seems to decrease dramatically as the novel progresses, regaining momentum only by the very end. Here, a cynical analyst might ask, is this attributable to *Middlemarch's* less interesting middle sections, or are these critics, especially in today's world of attention deficit disorders, simply skipping over the middle sections in order to reach the end?

Of course, there are a few reasons why the cynical theory is unlikely. One of the

most obvious is that nearly every chapter in the novel is quoted from at least once, with the exception of Chapter 44, from which no quotations were identified. (Let Chapter 44 serve as a challenge for early-career Eliot scholars looking to make a unique mark in the field!) Another is that literary criticism, despite being predominantly single-author, is an inherently social activity: critics respond to those that have published before them, and their choices of passages to cite and discuss are not determined entirely by their idiosyncratic interests, but are heavily conditioned by previous critical trends.

By far the most quoted chapter, according to this analysis, is Chapter 20, which describes Dorothea's first impressions of Rome, and from which the passage that begins this article was excerpted. That passage is the culmination of a few paragraphs of rich, colorful descriptions of Rome and its artistic treasures, sections which contrast sharply with much of the stark provincial English drama of the previous chapters. It is also a chapter full of abstract concepts like *life, death,* the *world,* and *soul*.

In sharp contrast to Chapter 20 is the infrequently-quoted Chapter 86. Despite the overall interest in the end of the novel, for instance, this penultimate section is cited almost not at all. This might be explained in part by its very short length. More importantly, however, it consists almost entirely of dialogue, and adds nothing new to the plot development. It might be seen as a kind of dénouement between the oft-quoted climactic Chapter 80 and the heavily-quoted Finale. Most notably, it contains virtually no aphoristic language—no summarizing or abstracting commentary from the narrator.

When we divide these quotations into the decades in which their articles were published, a somewhat different picture emerges. Although our corpus contains critical articles from as early as the 1860s, we begin here with the 1950s, since this is the decade in which

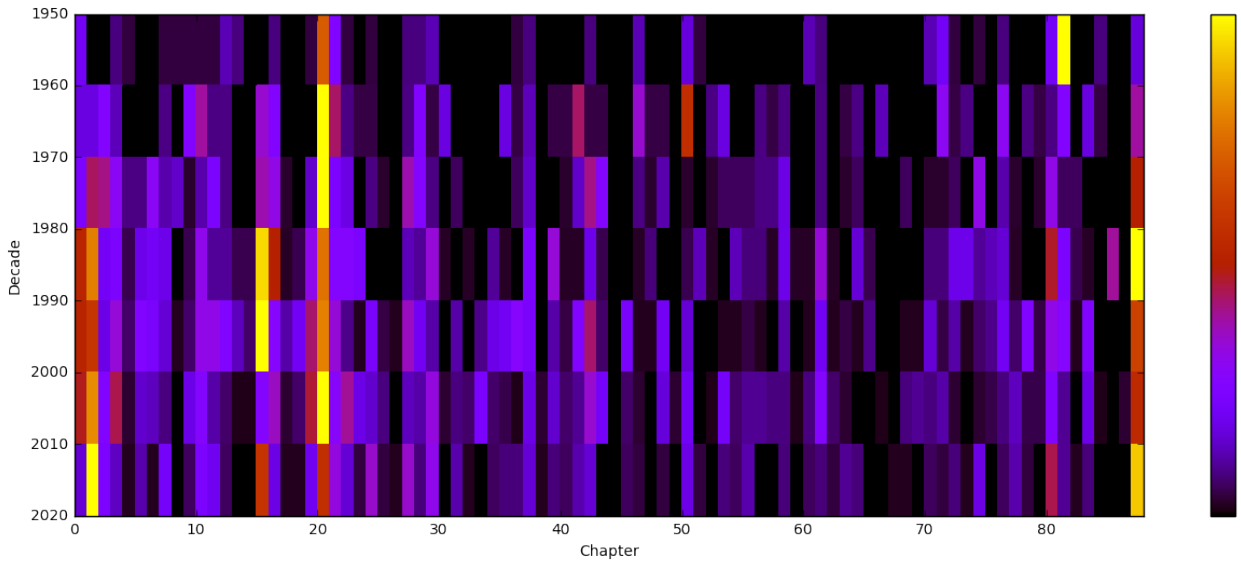

Figure 4: Diachronic Analysis of Middlemarch Citations, by Chapter

the number of articles first surpasses one hundred. Figure 4 shows this diachronic analysis. In this chart, the color represents the number of quotations relative to that decade, with black the least-cited chapter per decade, and yellow the most. Chapter 20 remains fairly constant through the latter 20th century, but critical attention seems to shift in other areas. Chapter 81, for instance, which was the most-cited chapter in the 1950s, quickly falls out of favor with critics, who become more interested in the beginning and end of the novel. In the 1980s and 90s, Chapter 15 becomes the most-cited chapter in *Middlemarch*, but falls out of fashion thereafter. In fact, with the exception of Chapter 20, the trend here, if one reads the history of each chapter in this figure vertically, is one of coming into fashion and passing out of it.

As mentioned above, one of the most-represented journals in this corpus is the specialist journal, *George Eliot - George Henry Lewes Studies*. We hypothesized that this journal would show different sorts of citations than more generalist journals such as *Victorian Stud-*

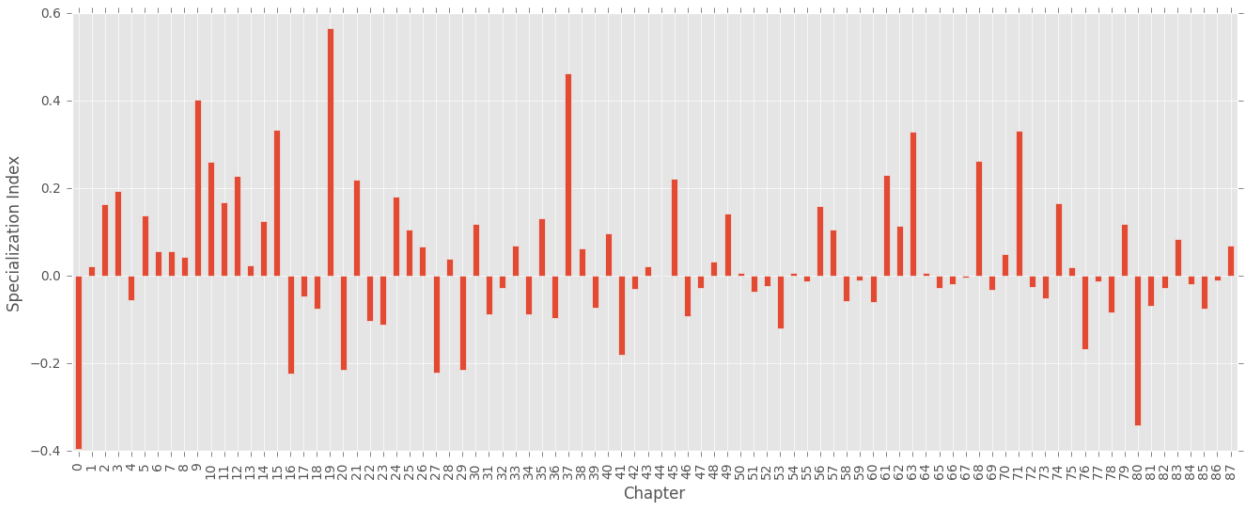

Figure 5: Comparison of GE-GHLS with Generalist Journals

ies. Figure 5 shows that comparison—a normalized factor of its relative citation frequency between these two categories. Chapters with positive specialization indices are cited more frequently in the specialist journal, while those with negative indices are cited more frequently in generalist journals. What is most striking about this picture is the divergence of these two categories: with the exception of the ten or twenty chapters with indices close to zero, most of these chapters are polarized in one direction or another. The novel's prelude is chiefly the domain of generalist journals, while the rest of the first quarter is primarily the domain of the specialist journal. Chapter 20, the most-cited chapter overall, is cited more in generalist journals, perhaps because it is the most famous section in the novel. It is perhaps for this reason that it doesn't seem to be of much interest to specialists, who are more concerned with the less popular surrounding sections.

3. The Language of Quotations

What makes a passage quotable? To help answer this question, we performed a linguistic analysis of quoted and unquoted text. We began by extracting quoted passages from our annotated edition, and binning them according to how frequently they were quoted. From there, we tagged all words by part of speech using the SpaCy tagger, and measured the frequency of these parts of speech in these bins. Figure 6 shows how many more times a given part of speech appears in quoted text than unquoted text—(a) showing the difference between unquoted text and quoted text of all levels, and (b) showing the same difference, but with highly quoted text—text quoted over twenty-five times. These parts of speech follow naming conventions set by the Penn Treebank (Marcus, Marcinkiewicz, and Santorini 1993, 317). Parts of speech with positive scores appear more frequently in quoted text, while those with negative scores appear more frequently in unquoted text.

The most distinctive part of speech for quoted text—not represented in Figure 6 because it was made the chart unreadable—is possessive wh-pronouns, denoted "WP$." These are almost twice as common in quoted text, and twelve times as common in highly quoted text. In *Middlemarch*, these correspond entirely to the word *whose.* These are almost all descriptions of Dorothea, such as the narrator's often-quoted depiction of her in Rome, as "a breathing blooming girl, whose form, not shamed by the Ariadne, was clad in Quakerish gray drapery," "a girl whose ardent nature turned all her small allowance of knowledge into principles," and "whose quick emotions gave the most abstract things the quality of a pleasure or a pain" (Eliot 2004, 176, 180). Perhaps critics frequently quote these passages as if they are definitions, keys to understanding Dorothea's character.

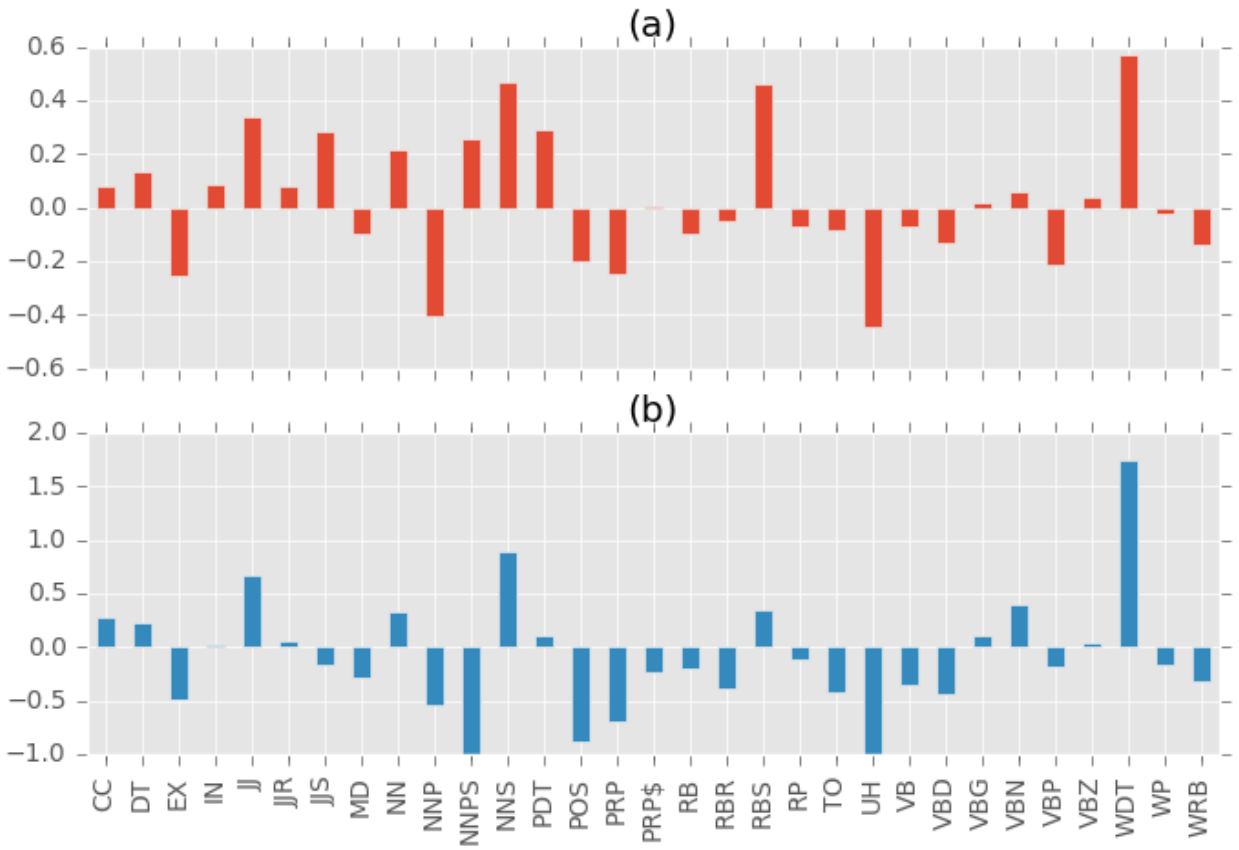

Figure 6: POS Comparison

Other frequent parts of speech in quoted text include wh-determiners, denoted "WDT," and encompassing the words *which, that,* and *whatever.* These markers often denote quasi-definitive description, as well, as in Dorothea's "insistence on regulating life according to notions which might cause a wary man to hesitate," and her "white beaver bonnet which made a sort of halo to her face" (Eliot 2004, 35, 176). *That,* too, appears in these definitive contexts, often accompanied by *all,* such as Dorothea's characterization of Rome as a place "where all that was living and warm-blooded seemed sunk in the deep degeneracy of a superstition," and the narrator's statement of focus, "that all the light I can command must be concentrated on this particular web" (Eliot 2004, 180, 137).

Also distinctive to quoted text are plural nouns, denoted "NNS," which appear about 1.5 times more often in quoted than unquoted text, and twice as often in highly quoted text. These plurals are mostly abstractions—mental processes such as *notions*, *interpretations*, *thoughts*, and *emotions*; scientific metaphors such as *atoms*, *vortices*, *conditions*, and *experiments*; but most notably, pathway-metaphors of life, such as *vistas*, *thoroughfares*, and *pathways*[3].

Curiously absent in quoted text are plural proper nouns, denoted "NNPS." These include family names such as *Raffles*, *Farebrothers*, and *Casaubons*, words that our text matcher never identifies in our critical works. Only a total of twenty-six plural proper nouns appear in the corpus of quoted text, most of which are names for categories or large groups of people, such as *Saints*, *Sages*, *Lords*, *Christians*, and *Israelites.* This distinction—between individuals and abstract concepts, is one which is most apparent in the distinctive-word analysis of quoted text.

[3]A full list of these may be found in the quoted speech notebook on the project code repository.

Our final analysis of quoted text measures the difference of relative word proportions between quotes and non-quotes. These words were first lemmatized, collapsing sets of related tokens such as *happy, happier,* and *happiest* to just *happy,* and forms such as *are* and *is* to *be.* The lemmas most distinctive of quoted text are, in this order: *life*, *woman*, *like*, *love*, *world*, *consciousness*, and *soul.* The list of the top twenty distinctive words includes more of these abstractions, as well as markers of magnitude, like *great*, *large* and *small*, markers of subjectivity, such as *mind*, *heart*, *self*, and *inward*, and words suggestive of wide scopes, such as *people*, *history*, *human*, and *nature.* These categories—magnitude, subjectivity, and abstraction—might be said to best characterize quoted text in *Middlemarch.*

The lemmas with the lowest scores, and thus most uncharacteristic of quoted text are, in this order, *say*, *Mr.*, *Lydgate*, *Fred*, *Bulstrode*, *Mary*, and *Rosamond*—titles and names of minor characters. It perhaps should be unsurprising that *say* is the lemma most uncharacteristic of quotes, since critics often paraphrase narratorial speech attributions. Rather than constructing a nested quotation, like " 'How very beautiful these gems are!' said Dorothea," critics usually quote the character directly, as in "this leads Dorothea to exclaim, 'How very beautiful these gems are!' "

Antoine Compagnon's theory of quotation, expounded in *la Seconde main*, is that it mediates between the acts of reading and writing (1979, 34). In fact, he argues that the textual practices of reading and writing are indistinguishable from the process of citation, which is one of selection and inscription. A writer first reads, underlines what he reads, and then rewrites ("écrire," he says, echoing Foucault, "c'est toujours récrire"). Underlining, therefore, is a kind of translation, from the dialect of the text into one's idiolect: "Le soulignement marque une étape dans la lecture, il est un geste récurrent qui paraphe, qui

surcharge le text de ma propre trace. Je m'introduis entre les lignes … je déchire les fibres du papier, je souille et dégrade un object: je le fais mien" (20). We might be able to explain the abstractions of quoted text, then, as mediations between the life of a reader and the lives of the novel's characters. Put differently, *life*, the lemma and the concept, is that which mediates between the world of the novel and the reader's inner world.

These observations of quoted text might lead us, ultimately, to a theory of reading. Do we, not only as critics and readers of *Middlemarch*, but as readers, have sharper memories for passages that speak to totalities of human experience, namely, to *life*, the *world*, and to *consciousness*? Do these abstractions facilitate intersubjectivities that allow us to access the otherwise distant worlds of fictional characters? Furthermore, are we selfishly motivated as readers, reading fictional works as if they are sacred texts, looking for instructions that will help us live our lives? Although only one text of many, *Middlemarch*, and the analyses of its readings discussed here, might provide the beginnings of answers to these questions.

References

Bär, Daniel, Torsten Zesch, and Iryna Gurevych. 2012. "Text Reuse Detection Using a Composition of Text Similarity Measures." *Proceedings of COLING 2012* 1: 167–84. https://www.cdc.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2012/COLING_2012_DaB_published.pdf.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. http://dl.acm.org/citation.cfm?id=2133826.

Compagnon, Antoine. 1979. *La Seconde Main : Ou, Le Travail de La Citation*. Paris:

Seuil.

Eliot, George. 1873. *Wise, Witty, and Tender Sayings in Prose and Verse: Selected from the Works of George Eliot*. W. Blackwood.

———. 1878. *The George Eliot Birthday Book*. Edinburgh: William Blackwood & Sons.

———. 2004. *Middlemarch*. Edited by Gregory Maertz. Peterborough, ON: Broadview Press.

Goldstone, Andrew, and Ted Underwood. 2014. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45 (3): 359–84. doi:10.1353/nlh.2014.0025.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. "Building a Large Annotated Corpus of English: The Penn Treebank." *Comput. Linguist.* 19 (2): 313–30. http://dl.acm.org/citation.cfm?id=972470.972475.

Moore, Ella Adams. 1903. *A Moment Each Day with George Eliot: A Quotation for Every Day in the Year Selected from the Works of George Eliot*. Madison book Company.

Price, Leah. 2003. *Anthology and the Rise of the Novel*. Cambridge, UK: Cambridge University Press.

Smith, D. A., R. Cordell, and E. M. Dillon. 2013. "Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers." In *2013 IEEE International Conference on Big Data*, 86–94. doi:10.1109/BigData.2013.6691675.

Tenen, Dennis. 2017. "Digital Displacement." In *Futures of Comparative Literature*, edited by Ursula K. Heise, Dudley Andrew, Alexander Beecroft, Jessica Berman, David Damrosch, Guillermina De Ferrari, César Domínguez, Barbara Harlow, and Eric Hayot.

London: Routledge.