How Scholars Quote from Primary Texts: a Case for Domain-Specific Text Reuse Detection

# # Abstract [100 words]

Using text-reuse detection, we develop a tool for analyzing quotations from primary texts in a body of scholarly writing. While text reuse detection has increasingly been gaining traction with computational literary studies and adjacent computational fields, less attention has been paid to the particular problem that algorithmic detection of quotations from primary sources poses. Our tool and subsequent analysis of a collection of JSTOR-held scholarly writing demonstrates the need for domain-specific text reuse detection. Outlining the importance of a domain-specific tool, we offer a methodology for tracing not only how and where a text has been reused, but a more granular analysis of what parts have been used over time. Bringing text-reuse methods together with rich bibliographic metadata, we showcase the strengths of this local, more domain-specific method for identifying quotations.

# # Introduction

Quotations, acts of plagiarism, instances of repeating gene sequences–these represent just a small fraction of the different kinds of objects that a researcher might be after when they look for instances of text reuse. Within the context of computational scholarship, text-reuse detection––or the algorithmic identification of passages that appear in two sets of text––has benefited from its ecumenical approach to texts. The ability to trace the reappearance of a passage or part text in another useful for understanding many humanistic domains that cite and use primary sources, from law to historiography, as well as many realms in genomics that deal with text in the form of DNA. Such methods, it would seem to offer a wide range of disciplines the ability to trace instances of textual borrowing or reuse.

Despite its universal promise, text reuse detection as a field often answers a much narrower kind of research question: is there a possibility that portions of one text appear in another (often for the purposes of determining the possibility plagiarism or textual borrowing). But this assumption, and the assumption that methods can be used interchangeably might be used for any text or corpus, are not well suited to all forms of textual that scholars might ask. In our case, we take literary studies as our case study, a field where the long-standing central disciplinary norm of close reading and textual exegesis gives quotations from primary texts a particular role in literary argumentation. In this case, detecting what counts as a true quotation from a primary text requires a more local and domain-specific form of text-reuse detection with a shift in frames. Instead of asking *if* there’s a possibility that one text includes text, our focus shifts to what parts of a text (often a primary text) appear in another. While literary scholarship as a case study, the method we outline one method for local and domain specific text reuse could be adapted and modified for other disciplines and their conventions around quotation of primary texts. Our central claim is that the questions that text reuse detection helps answer are broader than the current, more domain-independent methods can address.

# # Background

Computational text reuse detection has benefited from development in several specific domains. Plagiarism detection, a special use-case of text reuse detection, was primarily developed to solve, as one 1981 work puts it, "the problem of plagiarism in programming assignments by students in computer science courses" [@donaldson1981]. A computer science programming assignment, however, differs greatly from literary scholarship. For one, whitespace and punctuation are significant. For another, there are cases where quotation is disqualified as plagiarism, but remains a case of text reuse. In many cases, plagiarism detection software uses some form of document similarity measure, since the assumption is that a plagiarized assignment is copied wholesale from another document. There have been several useful overviews of the state of the plagiarism detection problem, a long-standing problem in natural language processing [@lukashenko2007].

Text reuse detection has also been used in genomics and bioinformatics generally. The task of gene sequencing has led to the use of sequence alignment algorithms, which have undergone various developments over the years [@heng2010]. Given two strings of nucleobases cytosine [C], guanine [G], adenine [A] or thymine [T], such as "ACCGATTAGA" the algorithm aligns them, such that the common nucleotides have identical indices.

ACCGATTAGA

ATTAGACGGACG

One of the most well-known algorithms is BLAST, the Basic Local Alignment Search Tool of 1990 [@altschul1990basic]. As with plagiarism detection, this problem differs greatly from ours, in that the possible elements are completely consistent and few in number (A|C|G|T). With quotation from literary sources, the elements to be sequenced potentially include every word in a language. Moreover, verb and noun forms may be changed in order to fit the scholar’s syntax ("he was 'descend[ing] down the staircase'") and, since human memory is involved, even misquotations sometimes make it into print.

Starting in the early 2000s, methods for studying text-reuse began to make their way from plagiarism detection and bio-informatics into computational linguistics. Some of the early methods were oriented around tracking the reuse of newswire stories from the AP or Reuters in other newspapers[@CloughMETERMEasuringTExt2002; @SeoLocaltextreuse2008; @BarTextReuseDetection2012] or the study of plagiarism communities[@KhritankovDiscoveringtextreuse2015], while others developed methods for detecting repeated sequences of text incidentally (as part of generating hyperlinks between different parts of the Google Books corpus.[@KolakGeneratingLinksMining2008]

Early examples of literary scholarship that used text-reuse detection were modeled after these gene-sequencing and plagiarism detection methods and, as a consequence, text-reuse methods were shaped around a particular problem: identifying *any* instances of potential similarity between two collections of texts. Scholars were chiefly interested in the potential evidence of literary borrowing, using BLAST algorithms to find evidence that eighteenth-century reference texts borrowed passages from Diderot's and D’Alembert’s encyclopedias,[@OlsenSomethingBorrowedSequence2011] or whether 19th-century French writers like Balzac and Gautier plagiarized from one another.[@GanasciaAutomaticdetectionreuses2014]

Text-reuse detection has, more recently, been seen as a method for identifying communities produced by the reprinting of texts, particularly in newspapers and periodicals. In their work on the *Viral Texts Project*, Ryan Cordell, David Smith, and Elizabeth Maddox Dillon have shown how reprinted poems, ads, and short news articles in 18th-and 19th-century American newspript can illuminate patterns in newspaper syndication.[Project [@SmithInfectioustextsModeling2013; @SmithDetectingModelingLocal2014] Their work has paved the way for other scholarship on reprinting, including newspapers in other languages,[@VesantoApplyingBLASTText2017; and @VesantoSystemIdentifyingExploring2017 both apply the BLAST algorithm to Finnish newspapers in the 18th-19th century] and towards different aims, like tracing the recirculation networks of newswire copy in the corpora of 21st-century news articles from the US and the UK.[@NichollsDetectingTextualReuse2019] In a similar vein as text-reuse in literary studies, computational linguistics have used similar methods in examining how the text of a given policy document is reprinted in other bills and legislative texts.[@LinderTextPolicyMeasuring2020]Here, while occasional attention is paid to the particular texts that “go viral,” the focus is not on particular passages but on the networks of publishers, syndicators, and editors that such acts of reprinting reveal.

When computational scholars do focus on quotation, their studies of quotation typically focus on quotation of passages from canonical primary texts (often, texts from Greek and Roman antiquity or the Hebrew or Christian Bible) in other primary texts.[@GessnerBiblicalintertextualitydigital2013; cite Jonathan’s work on Bible quotations] Lincoln Mullen, in *America’s Public Bible*, focuses on quotations from the Old and New Testaments of the Bible in a corpus of American 18th- and 19th-century newspapers.[@LincolnAmericaPublicBible; @QuotationFinderAmerica2022] Marco Büchler’s TRACER project offers something of a hybrid between the study of particular quotations and their circulation: TRACER focuses on reprinting of Classical Greek authors within a corpus of Greek historian’s reference texts and takes frequent quotation as a metric for a work’s “influence.”[@BuchlerMeasuringInfluenceWork2013; @KokkinakisDetectingReuseBiblical2016; see also Frederik Arnold's “Lotte” (later renamed “Quid”) framework[@ArnoldLottev12022; @ArnoldQuid2022; @ArnoldLotteAnnetteFramework] Detecting quotations within primary texts poses particular challenges: what edition and translation of the Bible to use, or how to deal with slight misquotations[@DuhaimeTextualReuseEighteenth2016; @RoeDiggingECCOIdentifying2016] Recent work has also begun to extend text-reuse methods to non-Latin scripts[@SturgeonUnsupervisedidentificationtext2018; @Budakdirectphonologydphon2021; @Budakdphon2022]

A related but distinct area of scholarship to the study of text reuse is the study of \*citations\*, which often goes by the name “bibliometrics.” Whereas quotation involves the verbatim replication of some portion of the source text, citation involves only a reference. In the case of scholarly writings, this is usually a bibliographical reference formatted according to a citation style determined by the publisher. For example, to cite Foucault’s concept of the “author-function” or the article “What is An Author?” from which it derives is distinct from quoting his specific formulations. Citations have particular significance for scholars within the contemporary university, where they are a widely-used metric that determines career opportunities. Many decades of research have demonstrated that the distribution of citations in academia replicates and reinforces broader inequalities related to social categories like gender and race. [See e.g. @FerberCitationsAreThey1986; @EarhartCitationalPoliticsQuantifying2021; @CiteBlackWomenCollectiveCiteBlackWomen] Focused as they are on the social status of the cited works’ authors, these studies don’t typically examine the smaller scale of which portions of a work are cited. [See @RomanelloExploringCitationNetworks2016 for an exception, which also analyzes the book and line numbers included in citations of classical texts like Vergil’s \*Georgics\*.] Methodologically, all these studies function not by detecting text reuse but by tallying bibliographical references, which in scholarly writing have a relatively consistent form (e.g. in-line citations; footnotes; endnotes; list of works cited). In contrast, our method detects instances of text reuse, whether accompanied by a bibliographical reference or not, and doesn’t encompass bare citations without quotation. While bare citation is the norm in certain natural sciences, across the humanities and in literary studies in particular, quoting from and citing a source are both options, whose distinctions we hope to understand better.[^01]

[^01]: Literary scholars have also studied quotation practices using various non-digital methods. Studies focusing on the presence of quotations in literary works include @MeyerPoeticsQuotationEuropean2015; @CompagnonSecondeMainOu1979; @PrinsVictorianSappho1999; @BuurmaEpigraphsMatter2012; and @HackReapingSomethingNew2016. Studies of how portions of literary texts have circulated in the form of quotations include @GarberQuotationMarks1999; @PriceAnthologyRiseNovel2000; @DamesNotCloseReading2010a; and @MoleWhatVictoriansMade2017. Closest to our own project, there have been several recent (non-quantitative) studies examining how literary scholars quote: @AuyoungWhatWeMean2020 and @KramnickCriticismTruth2020.

Our text-matcher builds on these existing bodies of work by developing methods geared towards a narrower domain: detecting quotations from primary texts used in scholarly writing. Unlike studies of text-reuse grounded in plagiarism detection or in networks of reprinting—where the threshold for determining whether a piece of text might be reproduced from another is lower in order to capture as many possible instances––our methodology is designed more conservatively in order to ensure that a given string of text is quoted from the text in question. In focusing on the quotation of primary texts within scholarship, we build a tool with the aim of identifying quotations in corpora (scholarly writings) that have more standard conventions for direct quotation. In doing so, we take inspiration from several recent digital projects studying scholarly quotation patterns in aggregate. JSTOR Labs and Derek Miller have analyzed quotations from literary texts in scholarly writings: both projects present interactive web visualizations for exploring what lines have been most quoted within Shakespeare’s corpus.[@HumphreysHowJSTORLabs2017;@MillerQuoteNotQuote] Our text-matcher has been used in the study of other corpora, most notably [Our text matcher has been used in @PiperMeasuringUnreading2020 to study quotations of Goethe’s corpus; related work on this topic has been done by the Stanford Lit Lab at MLA 2016]

^[While we were conducting these experiments, and presenting our initial findings at Digital Humanities 2017, we learned that the Stanford Literary Lab was working on a strikingly similar problem: text matching between a large corpus of literary texts, and a corpus of historical book reviews and critical writings from the British Periodicals Online collection. The Stanford Literary Lab independently arrived at many of the same parameters we use for text matching, even using the same Python library. ]

# # Methods

We begin by pre-processing a text into a sequence of tokens. We strip out punctuation and extraneous whitespace, and lowercase the text. We also concatenate hyphenated words, in order to match against phrases which have been typographically hyphenated. Notably, this preprocessing removes quotation marks, which could potentially have been useful. But since automated parsers are not very accurate at identifying material inside quotation marks, and since our OCRed texts sometimes contain quotation marks which are not present in the original, we disregard them.

From there, we convert the words into stems using the Lancaster stemmer of the NLTK, which uses the Paice-Husk stemming algorithm.[cite:@nltkLancaster;@paice1990] We chose this stemmer, after evaluating several other options, for its lemma-like stems, which retain enough of the semantics of the original text, while allowing for the variation between verb forms which we expect from scholarly quotation. Finally, we group these stems into n-grams---trigrams by default---allowing us to alleviate the computational work required of the SequenceMatcher.

The core matching algorithm we use is the SequenceMatcher from the Python library ~difflib~ [cite:@peters\_difflib\_2016], which adapts Ratcliff/Obershelp Pattern Recognition, or "gestalt pattern matching" [cite:@ratcliff1988pattern]. This computes the string similarity $D\_{ro}$ of strings $S\_1$ and $S\_2$ according to their matching tokens $K\_m$.

$$ D\_{ro} = \frac{2K\_m}{|S\_1|+|S\_2|} $$

This computes the initial, or core matches. By default, we search for a core match of length three, `--threshold=3` in the command line interface. This amounts to three overlapping trigrams---a total of five words. Thus, we define the minimal quotation as five identical stems. We arrive at this set of defaults after a long period of experimentation: too little of a threshold, we discovered, results in too many false positives, since there are many common trigram and 4-gram sequences which are features of the English language, rather than evidence of quotation. It must be emphasized that this core matching process must involve identical stems---given the origins of the Ratcliff/Obershelp algorithm in bioinformatics, we sacrifice this small amount of fuzzy matching to attain an initial performance boost.

However, the initial threshold only constitutes the core of the matching operation. From there, we perform two, additional functions, both of which are much slower, computationally. The first is to extend the match to contiguous words which may also be part of the same quotation. To do that, we look backwards and forwards from the boundaries of our initial match, and compare the Levenschtein edit distance ratios of the words at these boundaries. Levenschtein distance, or edit distance, describes the number of insertions, deletions, or substitutions required to edit one string into another. [cite:@levenshtein1966binary] When adjusted for the number of letters in each word, we can approximate a word's morphological similarity to another. So long as two words have an edit ratio of 0.4 or below, we consider them part of the same quotation. This allows us to handle differences in American and British spelling, as well as some OCR errors. Considering these example edit ratios:

| Word A | Word B | Edit ratio |

|---------+----------+------------|

| color | colour | 0.1818 |

| theater | theatre | 0.2857 |

| day | today | 0.5 |

| foobar | foo56bar | 0.2857 |

The edit ratios of the words with divergent British and American spellings are below the threshold and are considered matches; the similar words /day/ and /today/ are above the threshold and aren't considered matches. The word with OCR errors, /foo56bar/, is still considered a match.

Even with this forgiving algorithm, however, we noticed that matches were still breaking on certain words, when the number of OCR errors surpassed this threshold, or when the quoting text was interrupted by paratextual features, like running headers or page numbers. To mitigate these issues, we heal neighboring matches. If two matches are within eight tokens of each other, we concatenate them into the same match. This allows us to match strings across page breaks, matching, for instance, a string such as "hearing the grass grow and the squirrel's heart beat" with "hearing the grass GEORGE ELIOT--GEORGE HENRY LEWES STUDIES 54 and the squirrel's heart beat." Without this healing of neighboring matches, each individual match would be smaller than the minimum match size.

Analyzing quotations identified with our matcher requires good metadata. It was important to use that we were able to analyzing not only where in the text a particular quotation appeared (determined through a quick set of text processing to determine the index characters of each chapter in our sample text), but how quotation patterns varied over time within our corpus (in our case, a collection of 20th and 21st century scholarship from JSTOR). Extracting the dates from the metadata, we developed a method for diachronic analysis: analyzing what passages were quoted in different moments over time. JSTOR provided metadata for us that included both date of publication as well as journal of publication. This diachronic analysis allowed for more granular analysis of quotation patterns.

# # Results and discussion

What our results show is the importance of domain specific text reuse detection methods. Our algorithm and workflow is designed for a specific case and tuned with hyperparameters that are sensitive to the nature of a quotation in a literary text, offering a more nuanced method for humanities-specific applications.

Many of the modifications of our algorithm are designed to capture particular ways that literary scholars quote from primary texts. Take, for instance, this quotation that our healing of neighboring matches was able to correctly identify:

> PASSAGE IN MIDDLEMARCH: Bulstrode's sickly body, shattered by the agitations he had gone through since the last evening

> PASSAGE IN SCHOLARLY ARTICLE: Bulstrode's "sickly body" is "shattered by the agitations he had gone through since the last evening"[@CarpenterMEDICALCOSMOPOLITANISMMIDDLEMARCH2010, 521]

Using healing methods, we’re able to account for some of the ways that literary critics “knit” direct quotations from other texts into their own modifying the direct quotation to fit the syntax, style, mood, and tense of their own sentences.[@KramnickCriticismTruth2020, 223] In other instances, our fuzzy matching parameters and our healing were able to heal across instance of hyphenation (“ac-commodate” is correctly identified as one word).[@Millet] or, in this case, account for the match even text that was not directly included in the quotation but had been naturalized into the critic’s own language:

> PASSAGE IN MIDDLEMARCH: "growing good of the world is partly dependent on unhistoric acts; and that things are not so ill with you and me as they might have been, is half owing to the number who lived

faithfully a hidden life, and rest in unvisited tombs.

> PASSAGE IN SCHOLARLY ARTICLE: The narrator asserts that "the growing good of the world is partly dependent on unhistoric acts." If things are "not so ill ... as they might have been," Eliot concludes, it is "half owing to the number who lived faithfully a hidden life, and rest in unvisited tombs" (MM, 613).

Our text-matching algorithm makes intentionally conservative matches in order to avoid mis-identifying a string of text as a quotation. This means that we have relatively high precision (100% of the quotations our text matcher identified in a random, human-verified sample were in fact quotations from our source text). But this also come with some tradeoffs: we are only able to recall longer quotations 30% of the time. This is in part due to the hypersensitivity of our parameters: because our matcher is only set to determine a match if three matching overlapping tri-grams are found, in practice, the smallest sequence of text that we can identify is a string of 5 words long. We’ve set this parameter high because even if we were to match shorter strings, there would be no guarantee that the matched text is, in fact, a string from *Middlemarch*, and not a commonly occurring 1-2-, 3-, or 4-gram. Even removing stopwords would not be a solution, since some of the quotations that critics quote from the novel, like the the use of the phrase “you and me” in an epigraph––contain only stopwords. Our analysis shows that detecting quotations under 5 words in length is a more complicated task than simply solving it might imply. Instead, as this method advocates, detecting smaller n-grams, that cannot be solved by a “better” matching algorithm, since such an algorithm presumes the uniqueness of the quotation being matched––-the fact that that quote would appear only in that text––rather than acknowledging that commonly recurring 1- 2-, 3-, 4-grams constitute part of the text of a work that a literary critic might quote in the process of close reading.

As previously stated, the experiment for which we developed text-matcher was an analysis of critical quotations of George Eliot's novel \*Middlemarch\*. We were interested in knowing how \*Middlemarch\* was quoted over time, and whether there were patterns to these quotations. With the gracious assistance of our friends at JSTOR Labs, we were able to obtain full texts of over 4,000 critical articles which contain the word "Middlemarch." (The specificity of this novel's title helped us greatly in this query.)

Since text-matcher keeps track of the locations of matches, we are able to answer a number of useful questions about trends in the critical quotation of \*Middlemarch\*. One of our early motivating research questions was, which parts of the novel are quoted the most, and which are quoted the least? After running the text-matcher on our corpus of JSTOR critical articles, we were able to determine the most frequently quoted passages, and categorize these according to chapter, the most meaningful narrative unit for our study. @Fig:byChapter shows these numbers of quotations, according to chapter. The most quoted chapter, in number of quotations, is Chapter 20. This holds true as measured by number of quoted words, as well.

![Number of Quotations of \*Middlemarch\* by chapter](byChapter.jpg){#fig:byChapter}

Next, we set out to discover the histories of these quotations, by comparing their numbers with the years in which the articles were published. @Fig:byChapterDiachronic shows the same breakdown of quotations per chapter, but shown across seven decades. What we found surprised us—with the exception of Chapter 20, which remains the perennial favorite among critics, many quoted passages came in and out of fashion in 20- to 30–year cycles. Certain chapters become more quoted for a time, and then gradually less quoted, until they are ignored altogether. One chapter in particular, the second-to-last, is cited hotly in the 1980s, and then never again.

![Number of Quotations of \*Middlemarch\* by chapter, 1950–2020](byChapter.jpg){#fig:byChapterDiachronic}

We discovered that Chapter 20 is the most-quoted chapter, due to large numbers of quotations of its fifth and sixth paragraphs. This led us to develop an even more granular visualization of \*Middlemarch\*'s critical quotations: we created an annotated edition of the novel, in which its passages are colored according to the relative number of critical quotations: darker colored passages are less-quoted, and lighter colored texts are more quoted. @Fig:annotated shows a screenshot from this annotated edition, showing the most-quoted passage in the novel: the sentence, "if we had a keen vision and feeling of all ordinary human life, 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."

^[The annotated edition is available here:

https://lit-mod-viz.github.io/middlemarch-critical-histories/annotated.html ]

![Annotated edition of \*Middlemarch\*, by numbers of quotations](byChapter.jpg){#fig:annotated}

Our inspiration for this annotated edition comes from JSTOR's \*Understanding\* series. At the outset of our project, JSTOR had provided editions of Shakespeare plays and the US Constitution, annotated according to the number of JSTOR articles. Now, there are many more texts available for exploration, even \*Middlemarch\*. But while JSTOR's editions provide paragraph-level counts of quotations, our analyses are accurate to the character level. This enabled us to find, for example, that of this most-quoted sentence above, its beginning, "if we had a," is quoted much less than the rest of the sentence.

[Keywords]

These findings led us to investigate whether there were any patterns to quoted phrases: is it just their poetic lyricism? We thus tallied all the words that appear in quotations, weighted them according to numbers of quotations, and compared these word frequencies to the words that never appear in quotations. The result is a list of words that are quotable: if these words appear in /Middlemarch/, they are likely to be quoted by critics.

Quotable words include, in order of quotability: \*life, like, woman, dorothea, love, world, soul, consciousness, little, sort, deep, live, nature,\* and \*history\*. This list contains one character's name, Dorothea, which correlates with frequent quotations; the rest are chiefly abstractions. This leads us to posit whether critics are drawn to passages that contain claims---especially claims which deal with abstractions such as \*life, love,\* or \*soul\*---since they provide a easier entry into literary argumentation than, say, a bare description. We also wonder whether readers seek out abstractions as a way of making sense of the rest of the novel.

On the opposite end are non-quotable words: words that make a pass less likely to be quoted: \*say, Mr., Fred, Bulstrode, Lydgate, Mary, Garth, Celia, James,\* and \*not\*. The presence of the lemma \*say\* in this list indicates to us that dialogue indicators, like "said Bulstrode," are seldom quoted, but rather paraphrased. Otherwise, this reads like a list of minor characters, whom critics seem less interested in discussing.

It turns out that both of these lists of words are corroborated by the work of the Stanford Literary Lab, who find the same patterns with the British Periodicals Online corpus.

Conclusion

In this article, we’ve shown the importance of domain-specific studies of text reuse, using our case study in literary criticism of a single primary text (George Eliot’s Middlemarch) and JSTOR,. Our goal has been to provide a concrete pipeline and workflow for others to study quotations in primary sources and other corpora of secondary criticism, and to provide a clear pipeline for others using which is available for use on our GitHub repository. While we continue to refine our matching algorithm––in particular, with the aim of matching more 5-word but less than 12-word strings, our goal has been to illustrate some of the parameter tuning that comes with detecting quotations within a humanities field where quotations compose a significant portion of literary texts. While our primary case study has been in literary studies, our aim has been to demonstrate the value of local and context-specific methods for studying what parts of texts are quoted, and to offer pathways for further work on different corpora.